The antigenic evolution of human influenza A haemagglutinin

Thesis submitted for the degree of Doctor of Philosophy

The antigenic evolution of human influenza A haemagglutinin

William Dunbar Lees

Birkbeck, University of London
15th September, 2013

I, William Dunbar Lees, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

A detailed understanding of the B-cell response to influenza A haemagglutinin is key to the accurate matching of vaccines to seasonal strains, and may inform the development of broader spectrum vaccines. In this study, I develop techniques for predicting the location of the epitopes of protective antibodies by observing the physical locations of amino acid substitutions in human wild-type strains. By linking the understanding gained from this analysis with a large body of assay data, I present a model which can predict antigenic distance from HA1 amino acid sequences and which meets or exceeds the predictive power of previously developed models while retaining generality.

An interesting conclusion from the epitope analysis discussed above is that antibodies to the HA head bind in two regions. The antigenic evolution of influenza H3N2 is more punctuated than its genetic evolution. I propose that the dual regions might contribute to the punctuated nature of antigenic evolution, and explore this through the use of a simple simulation.

Stalk-binding antibodies to HA have attracted much interest in recent years: a number of broad-binding examples have been isolated, and the slower evolution of the stalk gives hope that these may provide broad protection against future strains. Stalk-binding neutralising antibodies to H3 are known to bind in two regions, and I use data from crystal studies to identify the constituent residues of these regions, which I term antigenic sites F and G, in a manner that is consistent with previous analyses of the constituent residues of HA1 antigenic sites A-E. I analyse the degree of conservation of residues in sites F and G, and conclude that there have been episodes of change in the H3 stalk which are consistent with antigenic evolution.

Acknowledgements

I would like to thank my supervisors Dr Adrian Shepherd and Prof. David Moss for their unstinting support, patience and encouragement, and for many interesting discussions over the course of this and my earlier work. Thanks also to my fellow student, Kevin Pentiah, for his insights and help.

I am very grateful to Dr John McCauley and Dr Rod Daniels of the UK WHO Collaborating Centre at the National Institute of Medical Research, who have been extremely generous with their time and have provided many insights. I would also like to thank the UK CC for publishing a large volume of assay data on a regular basis: this forms an invaluable historic record both for me and other researchers in the field.

Finally, thanks to my wife, Sue, for her unfailing support and understanding, and apologies for the many uncooked meals and absences.

Contents

List of Tables

List of Figures

Abbreviations

ELISA	Enzyme-linked immunosorbent assay
Fab	Antibody fragment, antigen binding
HA	Haemagglutinin
HA1, HA2	The two protein chains of haemagglutinin
HCDR	Antibody heavy chain complementarity determining region
HCV	Hepatitis C virus
HI	Haemagglutination inhibition
HIV	Human immunodeficiency virus
IgG	Immunoglobulin G
LAIV	Live attenuated influenza virus
LCDR	Antibody light chain complementarity determining region
MCC	Matthews correlation coefficient
MDCK	Madin-Darby canine kidney cell
MD	Molecular dynamics
MDS	Multi-dimensional scaling
ML	Maximum likelihood
NA	Neuraminidase
NCBI	US National centre for Biotechnology Information
RBS	Receptor Binding Site
SLAC	Single Most-Likely Ancestor Counting
TIV	Trivalent inactivated vaccine
WHO	World Health Organisation

1 Introduction

The global surveillance of influenza, mediated through the World Health Organisation (WHO), has yielded a substantial amount of sequence and antigenic information covering the evolution of the virus over the past 40 years. In this study I make use of that information to develop a deeper understanding of the specifics of antigenic evolution. The key objectives that underlie this work are:

- The prediction of antigenic change
- A greater understanding of the epitopes of protective antibodies to wild type strains
- An understanding of the mechanism that gives rise to the clustered antigenic development observed in H3N2 antigenic maps.

I first developed an approach for predicting epitopes of protective antibodies by observing the location of amino acid substitutions in successive dominant wild-type strains. An interesting result from this study was an indication that such epitopes group in two regions on the haemagglutinin monomer. Consideration of the more membrane-proximal of these two regions led to speculation that antigenic activity might occur not just in haemagglutinin's HA1 polypeptide chain but also in the HA2 chain, which is not normally considered to be evolving antigenically. My research coincided with an exciting period in which a number of human HA2-binding antibodies were isolated and studied in detail. I was therefore able to augment this computational study with experimental results from these studies, in order to form a hypothesis of antigenic evolution in HA2 that is consistent with rich experimental detail.

In the course of the study, I developed a number of computational approaches which can support influenza surveillance and the research of influenza and other viruses, in particular:

- Models to predict antigenic escape from sequence analysis, which can provide indicative results much more quickly than laboratory assays;
- The use of large bodies of sequence data in combination with structural information in order to identify evolutionary events that are close both in distance and in time;
- Visualisation tools that can rapidly and conveniently display information from an extensive sequence and antigenic database.

While a number of influenza sequence databases are publicly available, to the best of my knowledge there is no equivalent publicly available antigenic database.

As both sequencing and antigenic surveillance become ever more pervasive, computational techniques such as the above will become applicable to a wider set of organisms, and, correspondingly, their use will become increasingly necessary in order to draw conclusions from the large amount of available data. The amount of data available for influenza itself is continuing to increase, putting a strain on existing analytical methods and creating new opportunities for research. I discuss the implications for the future in the concluding chapter.

1.1 Influenza Structure

Influenza A is an enveloped negative-sense RNA virus (Figure 1.1). It has two surface glycoproteins: haemagglutinin (HA), and neuraminidase (NA). HA carries the receptor binding site (RBS), which attaches to sialic acid located on proteins on the surface of the host cell. NA is a sialidase which removes the sialic acid, allowing budding virions manufactured by the cell to escape. The viral envelope consists of a lipid membrane, given strength and shape by an inner 'matrix' protein, M1. A further matrix protein, M2, forms an ion channel through the envelope. The segmented genome consists of eight genes, each encapsulated in the nucleocapsid protein NP. Also present in the envelope are the proteins NEP, which mediates export of protein from the cell's nucleus during assembly; NS1, which has various roles in circumventing the cell's antiviral responses; and the proteins comprising the viral polymerase, PA, PB1, PB1-F2 and PB2. A final protein, N40, was discovered recently and is believed to have a role in viral replication (Wise et al., 2009). Detailed descriptions of structure and lifecycle may be found in Fields Virology (Palese et al., 2007) and Flint et al. (2009).

(this figure is not included in the public version)

Figure 1.1: Structure of the Influenza A virion, adapted from Flint et al. (2009). It is not clear whether the recently discovered protein N40 is present in the envelope or merely synthesized during viral replication. It is shown in this figure for completeness.

HA is a trimeric molecule in which each monomer is composed of two protein chains, HA1 and HA2 (Figure 1.2). The RBS is located in the globular head of HA1, towards the membrane-distal tip of the protein. The RBS binds to sialic acid at the terminal end of sugar chains on the surface of host cells. In humans, on the cells in question, most sialic acid present has an α-2,6 linkage, while in avian species, it mostly has an α-2,3 linkage. For an avian strain of influenza to reproduce efficiently in humans, adaptions to the RBS are required. Pigs have both receptors, and may present a bridge for such adaption (Klenk, Garten, and Matrosovich, 2011).

HA is the major antigenic component of the virus (Wright, Neumann, and Kawaoka, 2007): the primary mechanism through which human hosts gain immunity to an influenza A strain is through the development of antibodies which neutralise the virus by binding to HA in the vicinity of the RBS and hence block host cell attachment. In addition to its role in host cell attachment, HA also undergoes a conformational change during the process of endocytosis. The conformational change enables fusion of the endosomal membrane with the viral envelope, releasing viral RNA and protein into the cytoplasm (Wiley and Skehel, 1987). The virus can be neutralised by antibodies which interfere with the process of membrane fusion by binding to the stem in the neighbourhood of the fusion peptide. The action and prevalence of such antibodies is discussed in Chapter 5.

Influenza A HA and NA are divided into antigenic subtypes, determined by serology. Currently there are 17 subtypes of HA (1-17) and 9 of NA (1-9). Structurally and phylogenetically, the HA and NA subtypes both fall in to two groups. For HA, group 1 consists of subtypes H1, 2, 5, 6, 8, 9, 11, 12, 13, and 16, while group 2 consists of subtypes H3, 4, 7, 10, 14, and 15 (Skehel,

2009). The recently isolated H17 is thought to be closer to group 1 than group 2, but definitive phylogenetic analysis of this subtype has not completed (Tong et al., 2012).

Through the breeding of escape mutants to mouse monoclonal antibodies, it was established that neutralising antibodies to human H1 and H3 strains bound to defined regions in the HA head (Wiley, Wilson, and Skehel, 1981; Caton et al., 1982; Wiley and Skehel, 1987; Wilson and Cox, 1990). These are known as the *antigenic sites*. In H1, there are five sites, named Sa, Sb, Ca_1, Ca_2 and Cb. In H3, there are also five sites, named A-E. Bush and co-workers subsequently compiled a list of amino acid residues comprising the five antigenic sites in H3 (Bush et al., 1999), which I describe in greater detail in Section 5.6. This has been widely adopted in bioinformatics studies of H3 antigenicity and evolutionary pressure, and I shall describe it in this work as the 'canonical list' of H3 antigenic residues.

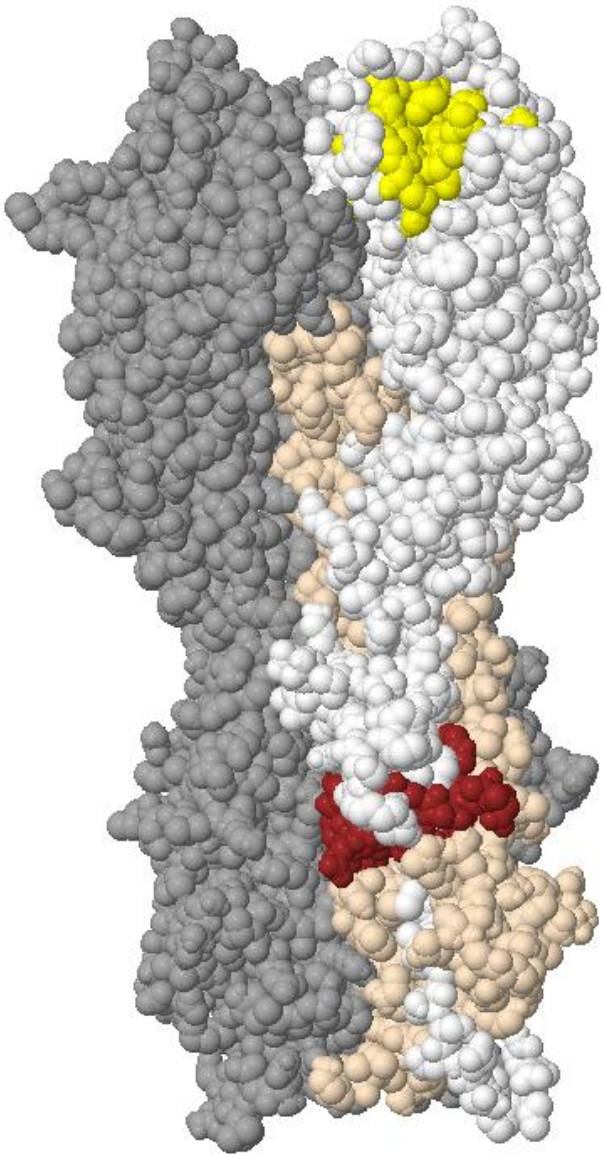


Figure 1.2: Spacefill view of H3 haemagglutinin, from the crystal structure derived by Sauter et al., (1992), rendered in Jmol. One monomer is shown in white (HA1) and light brown (HA2), with the RBS shaded yellow and the fusogenically active region, consisting of the N-terminal residues of HA2 shaded dark red. The other two monomers are shaded grey.

A subtype of influenza A is named according to its HA and NA subtypes, for example H3N2. The name of an influenza strain carries its type, host organism, place of isolation, strain number, and subtype, for example A/Human/Hong Kong/1/1968 (H3N2). The host designation is often dropped in works, such as this, where the host is clear from the subject and context.

1.1.1 Epidemics and Pandemics

In humans, influenza A HA undergoes rapid evolution under the selective pressure of neutralising antibodies. As a result of this evolution the virus can escape antibody action and re-infect hosts that gained immunity to previous strains. This process of *antigenic drift* leads to regular epidemics, infecting 5-15% of the population each year. Drift variants typically persist for 2-5 years. In the US, influenza epidemics are estimated to cause on average of 20,000 deaths per year (Wright, Neumann, and Kawaoka, 2007). Influenza epidemics occur in the winter season in temperate regions, and as a result this disease is known as seasonal influenza.

Influenza A infects a large variety of mammalian and avian species. Pigs and waterfowl in particular can act as reservoirs of different strains. Where a host becomes infected with multiple strains, the segmented genome lends itself to reassortment, occasionally resulting in the emergence of a novel strain. These novel strains, created by the process of *antigenic shift*, can cause pandemics in humans and other species. Since the beginning of the 20th century, there have been five human pandemics: 1918 (H1N1), 1957 (H2N2), 1968 (H3N2), 1977 (H1N1) and 2009 (H1N1). Of these, the most severe was the 1918 pandemic, which is estimated to have caused 25 million deaths in the first 12 months (Johnson and Mueller, 2002). Subsequent pandemics have been less severe. A definitive report on the 2009 pandemic is awaited from the WHO, but, in an initial study, global excess mortality in the first 12 months has been estimated at approximately 250,000 deaths, of which 80% were of individuals aged 65 or less (Dawood et al., 2012). In contrast to epidemic influenza, pandemic influenza is not seasonal.

Pandemics can result in the displacement of one circulating human subtype by another. The H1N1 subtype circulated from 1918 until the 1957 pandemic, when it was displaced by H2N2. H2N2 circulated until 1968, when it was displaced by H3N2. In 1977 the H1N1 subtype was reintroduced: 1977 strains were closely related to those circulating in the late 1950s. From 1977 to 2009, the H1N1 and H3N2 subtypes co-circulated. The 2009 pandemic strain is an H1N1 subtype but antigenically more similar to 1918 strains than those isolated more recently. It replaced previously circulating H1N1 strains and continues to co-circulate with H3N2. Its subtype is usually written as H1N1pdm or pH1N1, to distinguish it from the seasonal strains which it displaced.

1.1.2 Influenza Vaccination

Most vaccines today are based on the split-virus technique, licensed in 1968. In this technique, influenza virus is cultured in pathogen-free embryonated hen's eggs. HA and NA are split from the intact virion using detergents, and separated via ultracentrifugation. The seasonal vaccine consists of HA from three strains, selected to be representative of the currently circulating strains of A/H1N1, A/H3N2 and influenza B (Osterholm, Kelley, Manske, et al., 2012). This vaccine is known as the trivalent inactivated vaccine (TIV). It is delivered via intramuscular injection. Because the influenza B strains have split into two antigenically different clades, quadrivalent forms, containing two influenza B strains, are currently under development (Barr and Jelley, 2012).

An alternative vaccine, also available in many countries but much less widely used, uses cold-adapted live attenuated influenza virus (LAIV). The strains are cold-adapted so that they will infect the upper respiratory tract only. Again, the vaccine contains three strains as for TIV, and the strains are cultured in embryonated chicken eggs. This vaccine is administered as a nasal spray (Carter and Curran, 2011).

The strains used in seasonal vaccines must be regularly updated to allow for antigenic drift. The WHO conducts global surveillance of influenza strains through a network of national laboratories and regional collaborating centres. It issues recommendations for vaccine composition twice annually: in February for the Northern Hemisphere, and in September for the Southern Hemisphere. The timing allows approximately six months for national governments to review the recommendation and for manufacturers to create high-growth strains that will culture well in eggs, conduct trials, and ready stocks for the next season. In making the recommendation, the WHO's advisory panel has to exercise judgement to predict the antigenicity of strains that will be circulating in 6-12 months' time. They must balance the need to update a strain because of antigenic drift against possible manufacturing difficulties in culturing a new strain in eggs. The influenza vaccine is the only vaccine requiring annual review, update and administration to retain effectiveness.

The development timeframe for the seasonal vaccine provides a window during which the virus may develop unexpectedly: hence the vaccine is not always well matched against circulating strains. The time needed to develop a pandemic vaccine may be longer, because of the need to conduct additional safety trials in order to comply with national licensing requirements. Limited

supplies of the single vaccine against the H1N1 pandemic strain became available in the US in October 2009, and full stocks were available there in January 2010 – 5 and 8 months respectively after vaccine development started (Singleton et al., 2010).

A number of factors make it hard to determine vaccine effectiveness with accuracy. It is difficult to establish infection categorically. Some cases of influenza are asymptomatic. Where symptoms do occur, they are similar to those of other common infections such as rhinovirus, coronavirus and respiratory syncytial virus. Serology does not allow the effects of vaccination to be distinguished categorically from those of infection (McDonald and Andrews, 1955; Osterholm, Kelley, Manske, et al., 2012). Finally, the expected variation in vaccine effectiveness from year to year due to vaccine strain mismatch means that an extended trial is necessary for categorical results. A recent meta-analysis estimated the effectiveness of TIV in adults aged 18-65 at 59%. No studies meeting the criteria of the meta-analysis were available for other age groups. Effectiveness of LAIV in children aged 6 months to 7 years was estimated at 83%. Again no studies meeting the criteria were available for other age groups (Osterholm, Kelley, Sommer, et al., 2012). Even the higher of these two figures is below the likely level required for herd immunity, and it is notable that no figure is available for protection of the over 65 age group, which is the most vulnerable to seasonal infections.

A key goal in current influenza design is the elicitation of more effective, longer lasting antibodies. The stalk of HA is a potential target antigen, which we shall consider in this study (Pica and Palese, 2013).

This study is focussed on the humoral response of the human immune system. The response of the innate immune system and the cellular immune response are also important. The innate response alone is not able to clear influenza infection in experimental animals, but both CD4+ and CD8+ cells are able to clear infection independently (Wright, Neumann, and Kawaoka, 2007).

1.1.3 Determination of Antigenic Distance

Influenza surveillance requires a means of determining the extent to which antigenic drift has occurred. The assay generally used for this purpose in surveillance laboratories is the haemagglutination inhibition (HI) assay. The assay determines the extent to which serum taken from a host exposed to one strain of influenza is capable of neutralising the binding activity of a second strain (World Health Organization, 2011).

HA will bind multiply to erythrocytes in solution. In the well of a test tube or tray, this forms a dispersed red pattern. In the absence of such binding, the cells fall to the bottom of the well, forming a sharply defined 'button' shape. Antibodies binding to HA will inhibit this agglutination, and this inhibition forms the basis of the assay.

In a *haemagglutination* assay, two-fold serial dilutions of a viral sample are mixed with a specified amount of erythrocyte and placed in successive wells. The maximum dilution at which agglutination occurs is determined by observing the point at which the pattern in the well changes from diffuse red to button. This maximum dilution is known as the 'titre': a dilution of 1:128, for example, corresponds to a titre of 128.

For a *haemagglutination inhibition* assay, antiserum is collected by infecting a laboratory animal, typically a rabbit or ferret, and waiting for the infection to pass before collecting the serum. Reference antisera for well-known strains, such as vaccine strains, are kept in stock for use when required. In the assay, two-fold dilutions of antiserum are mixed with a fixed amount of viral sample and erythrocyte. The maximum dilution of antiserum that prevents agglutination is noted: this is known as the 'HI titre' and in this study we will refer to the titre between viral strain x and antiserum y as H_{xy}. We distinguish between homologous titres $H_{xx}$and heterologous titres H_{xy}.

The 'HI Distance' T_{xy} between strain x and antiserum y is defined as follows:

$$T_{xy} = \left(\frac{H_{yy}}{H_{xy}}\right)$$

(Burnet and Lush, 1940)

For the purposes of vaccine selection, an antibody is typically considered potent against a sample strain if T_{xy} is 4 or less (Schild et al., 1973).

Lapedes and Farber (2001) demonstrated that a 'shape space' of low dimensionality can be constructed in which antisera and antigens are treated as points, with the distance between them (the 'antigenic distance') being linearly related to the logarithm of the concentration ratio. The equation

$$D_{xy} = log_2\left(\frac{H_{yy}}{H_{xy}}\right)$$

defines the antigenic distance D_{xy}between the samples.

The terms HI distance (T_{xy}) and antigenic distance (D_{xy}) as defined above will be used in this study. The literature is not consistent in terminology, with some authors referring to T_{xy}as the antigenic distance, and others using the term for D_{xy}.

1.2 Outline

In Chapter 2, I summarise the sources of data used for this study, and provide an analysis of assay data quality.

Chapter 3 describes an analysis of H1N1 and H3N2 sequences directed towards understanding the binding location of human antibodies to wild-type strains. From this analysis, I derive a set of locations participating in H3 antigenic clusters, compare it with sets derived by other researchers through other means, and use it to develop a model that predicts the antigenic distance between two strains via amino acid sequence analysis.

An interesting conclusion from this work is that human antibodies to wild-type strains are found to bind in two regions of HA1. In Chapter 4 I develop a simple simulation model of antigenic distance in order to explore the influence that such two-region binding might exert on antigenic maps. I demonstrate that it is a possible cause of the clustered nature of the H3 antigenic map determined by Smith et al., (2004).

In Chapter 5, I examine possible roles of HA2 in the antigenic evolution of HA. I look first at its contribution to the antigenic clusters found in Chapter 3, identifying a limited role for HA2 in both H1 and H3. I discuss the identified stalk-binding epitopes of H1 and H3, and, using a database of HA2 sequences, make a case that the HA stalk, while more highly conserved than the globular head, is capable of antigenic evolution to escape antibody binding, and argue that evolution of the H3 stalk, as evidenced by substitutions and fixations, is compatible with such evolution.

In Chapter 6 I summarise some general conclusions from this work, and discuss future directions of research.

Examination of the ratio of non-silent to silent nucleotide mutations is a frequently used technique for identifying regions of a protein under positive or negative selective pressure. While not part of the main body of this work, in Appendix A I provide a brief comparison of this approach with the techniques used in the analysis above.

In Appendix B, I describe a website which brings together a number of analysis tools created for this work.

2 Assay and Sequence Data Quality

2.1 Analysis of HI data quality

2.1.1 Sources of Error

Because the biochemical properties of human influenza strains – such as binding specificity and avidity – have changed over time, experimental protocols have not remained constant. The Manual for the Laboratory Diagnosis and Virological Surveillance of Influenza (World Health Organization, 2011) defines current standard methods for surveillance laboratories, including a standard protocol for the HI assay. A number of issues relating to sources of experimental error are noted, as are the following points, which relate more to variation of biochemical and biological properties:

- Adaption of the viral strain to the culture medium, particularly in those strains cultured in eggs;
- Varying reactivity to specific red blood cells;
- High sensitivity of some viral strains to nonspecific inhibitors of haemagglutination.

These points are considered in more detail in following sections.

The repeatability of HI assays when conducted according to a defined standard operating procedure was studied in an H5N1-specific context by Noah et al. (2009), who concluded that reliable results could be obtained provided that experimental practice was carefully followed and appropriate controls were employed to allow issues to be identified and addressed. In particular, strict assay acceptance criteria must be observed, based on the results of assay controls. Assessment of haemagglutination must be based on standardised protocols, with variation between operators eliminated via training and assessment.

2.1.2 Viral Adaption in Culture

De Jong et al. (1988) observed that H3N2 viral specimens isolated from human subjects were heterogeneous: they contained a ‘major’ antigenic variant, and one or more ‘minor’ variants, with the minor variants being present in fractions between 10^{-1} and 10^{-3}. Culture of the sample in eggs or mammalian cells could be accompanied by changes in HI assay results as the proportion of variants adapted to the culture medium. Because samples cloned in mammalian cells were found to be stable upon further passage, while those cultured in eggs could take a

number of passages to stabilise, they argued that mammalian cultures should be preferred, and in recent years it has become customary in surveillance applications to culture strain samples in mammalian cells. Antisera, however, are typically raised from egg-cultured viral strains. Passage histories are not always provided in assay reports, but, for example, in the WHO UK Collaborating Centre's report of March 2000, approximately 50% of viral strain samples are egg passaged and 50% passaged in Madin-Darby canine kidney (MDCK) cells. In the WHO UK Collaborating Centre's report of February 2005, all viral strain samples are MDCK passaged.

The implications of the above for this work are, firstly, that both egg and mammalian cultured samples of a strain may be reported in the database, with potentially different antigenic results, and, secondly, that the amino acid sequence recorded for a strain may not match that of the variant that has been assayed.

2.1.3 Red Blood Cell Binding Avidity

Traditionally, the HI assay has employed chicken red blood cells, although turkey red blood cells are sometimes favoured as they provide a shorter settling time and inhibition patterns are clearer (World Health Organization, 2011). In the 1990s, however, a number of H3N2 strains emerged which were unable to agglutinate chicken red blood cells, and, subsequently, turkey red blood cells (Medeiros et al., 2001). Guinea pig red blood cells were able to agglutinate strains throughout this period, and are now recommended for the assay of all novel strains (World Health Organization, 2011). Published assay results through this period do not appear to have suffered quality problems as a result, although the variable avidity of red blood cells employed in successive assays, even when drawn from genetically similar animals, is a key source of variability between assays of the same viral strain against the same reference antiserum (private discussions with Dr John McCauley and Dr Rodney Daniels, UK WHO Collaborating Centre for influenza surveillance, November 2012). The red blood cell type used in preparing surveillance data is not routinely published.

A more serious problem developed in the early 2000s when it became apparent that a number of 'low reacting' MDCK cultured H3N2 strains would not develop high titres, even homologous titres against their egg-grown antisera, or could be made to do so only with great difficulty (see, for example, the summary of H3N2 activity in the introduction to the UK WHO Collaborating Centre interim report for March 2007). The cause of this problem was identified by Lin et al. (2010). Surprisingly, it results from an amino acid substitution in neuraminidase, which allows

neuraminidase to bind to red blood cells and hence cause agglutination in an antiserum-insensitive manner. Luckily this binding has proved to be oseltamivir-sensitive, and the addition of oseltamivir carboxylate to the assay restores the expected specificity.

In previous work (Lees, 2009), it proved necessary to exclude post-2003 assays with low-reacting homologous titres in order to obtain acceptable predictive results through these years. While that practice is now fully justified by the identified cause, it may be better to examine the neuraminidase sequence for the implicated substitution.

2.1.4 Non-Specific Inhibitors of Haemagglutination

Non-specific inhibitors of influenza virus haemagglutination are naturally occurring components of serum. Three types have been identified: the alpha class, which are sialyated glycoproteins that, while inhibiting haemagglutination, do not neutralise viral activity; the beta class, which are lectins that bind to mannose-rich glycoproteins on the head of HA and hence interfere sterically with the binding to red blood cells and inhibit viral activity; and the gamma class, which are similar to the alpha class, but also inhibit viral activity (Ryan-Poirier and Kawaoka, 1991; Anders, Hartley, and Jackson, 1990). While steps are taken in the WHO protocol to inactivate non-specific inhibitors, some residual activity may remain. This can affect the assay results of strains that are particularly sensitive to their presence, usually resulting in high titre values to multiple antisera.

2.2 Mathematical Treatment of HI Assay Results

The large volume of data in the HI assay database, and particularly the large number of measurements available for some reference strains and antisera, provides an opportunity to examine the repeatability of assay results.

2.2.1 Normalization of the Assay Data

Archetti and Horsfall (1950) introduced a ‘two-way’ measure of antigenic distance between two strains in order to normalise assay results against variations in experimental conditions. This measure is calculated from ‘cross titres’ of each strain against the other strain’s antiserum. Ndifon (2011) formalised this as follows.

Suppose that H_{ij} is the titre obtained with isolate *i* against antiserum raised from isolate *j*. Then

$$H_{ij} = A_j K_{ij} J_i$$

where A_j is the concentration of antibodies in the serum, $K_{ij}$is the average affinity of the antibodies for the virus, and J_i is a constant representing a range of factors such as the affinity between the virus and red blood cell.

We can see from this that the classical one-way distance,

$$D_{ij} = \log\left(\frac{H_{jj}}{H_{ij}}\right) = \log\left(\frac{K_{jj}J_j}{K_{ij}J_i}\right)$$

eliminates A_j but has a dependency on J_i and J_j. The two-way distance removes all unwanted dependencies:

$$D'_{ij} = \frac{1}{2}\log\left(\frac{H_{ii}H_{jj}}{H_{ij}H_{ji}}\right) = \frac{1}{2}\log\left(\frac{K_{ii}K_{jj}}{K_{ij}K_{ji}}\right)$$

Unfortunately it has not proved possible to determine a way of eliminating the dependence on J_i and J_j in the one-way measurements: instead, variation due to the underlying factors must be controlled as much as possible through experimental standardisation and treated as experimental error. On this basis, we may expect more variation in one-way assay results than in derived Archetti/Horsfall distances.

In this work, unless otherwise stated, the antigenic distance is defined as the log to the base 2 of the titre ratio. As the experimental procedure requires the titre ratio to be successively halved, using base 2 will generally yield integral distances.

2.2.2 Analysis of Errors

The HI assay data includes many multiple assays of the same strain/antiserum pair. Data to 2008 includes 39 pairs of H3N2 strains for which 10 or more assays are available. H1 data is more limited, but 6 or more assays are available for a total of 33 strains (Tables 2.1 and 2.2). For two-way Archetti/Horsfall data, where we have to combine two one-way measurements from the same HI table, the available data is more limited. In the same period, we can find 19 H3N2 strain pairs with 10 or more assays, and 16 H1 assays with 6 or more assays (Tables 2.3 and 2.4). In this section I shall consider the characteristics of the data sets listed in these four tables.

2.2.3 Distribution of Measurement Results

Figure 2.1 shows normal probability plots for the strain/antiserum or strain/strain pair for which there is most data. In the plots, the straight line represents the best fit of a normal distribution onto the data. The points should lie close to the line without any obvious skew or trend, and the mean and standard deviations calculated from the fit should be in agreement with the observed values. Mean and median values should be in near agreement.

Observed and calculated values of mean, standard deviation and median are provided in Table 2.5.

The relatively low number of measurements, and the low precision of the dataset (one-way measurements are typically integers between 1 and 8) limit the conclusions that can be drawn from this analysis, but the results are consistent with a normal distribution of results. Some estimated standard deviation values differ significantly in percentage terms from the calculated value, but the difference is small in absolute terms considering the precision of the measurements.

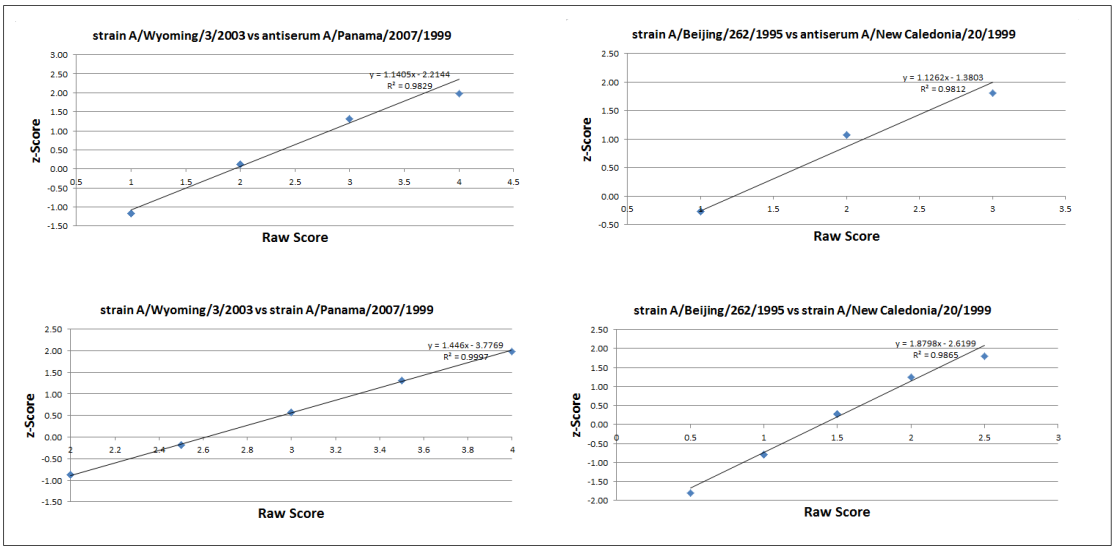


Figure 2.1: Normal Probability Plots for selected strain pairs from each dataset (H1N1/H3N2, one-way and two-way)

Strain	Antiserum	Assays	Log Distance	Std Dev	Std Err	95% conf +/-
A/Beijing/262/1995	A/New Caledonia/20/1999	14	1.29	0.61	0.16	0.33
A/New Caledonia/20/1999	A/Beijing/262/1995	14	1.93	0.62	0.16	0.33
A/New Caledonia/20/1999	A/Thessaloniki/24/2005	10	1.70	0.67	0.21	0.43
A/Thessaloniki/24/2005	A/New Caledonia/20/1999	9	0.33	0.50	0.17	0.33
A/Johannesburg/82/1996	A/Beijing/262/1995	8	3.38	0.92	0.32	0.65
A/Hong Kong/2652/2006	A/New Caledonia/20/1999	8	3.75	1.16	0.41	0.82
A/New Caledonia/20/1999	A/Egypt/96/2002	8	0.38	0.52	0.18	0.37
A/Egypt/96/2002	A/New Caledonia/20/1999	8	0.75	0.46	0.16	0.33
A/New Caledonia/20/1999	A/Hong Kong/2652/2006	8	3.38	0.52	0.18	0.37
A/New Caledonia/20/1999	A/Johannesburg/82/1996	8	6.50	0.76	0.27	0.53
A/Beijing/262/1995	A/Johannesburg/82/1996	8	5.38	0.92	0.32	0.65
A/Johannesburg/82/1996	A/New Caledonia/20/1999	8	3.88	0.83	0.30	0.59
A/Solomon Islands/3/2006	A/Hong Kong/2652/2006	7	1.14	0.69	0.26	0.52
A/New Caledonia/20/1999	A/Fujian/156/2000	7	1.29	0.76	0.29	0.57
A/Hong Kong/2652/2006	A/Solomon Islands/3/2006	7	1.29	0.95	0.36	0.72
A/Hawaii/15/2001	A/New Caledonia/20/1999	7	4.86	0.69	0.26	0.52
A/Fujian/156/2000	A/New Caledonia/20/1999	7	1.43	0.79	0.30	0.59
A/New Caledonia/20/1999	A/Hawaii/15/2001	7	1.29	0.76	0.29	0.57
A/Beijing/262/1995	A/Fujian/156/2000	7	2.14	0.69	0.26	0.52
A/New Caledonia/20/1999	A/Solomon Islands/3/2006	7	3.57	1.62	0.61	1.22
A/Hong Kong/2652/2006	A/Thessaloniki/24/2005	7	5.14	1.07	0.40	0.81
A/Fujian/156/2000	A/Beijing/262/1995	7	3.00	0.00	0.00	0.00
A/Beijing/262/1995	A/Hawaii/15/2001	6	2.33	1.03	0.42	0.84
A/Hawaii/15/2001	A/Beijing/262/1995	6	4.33	0.82	0.33	0.67
A/Chile/8885/2002	A/New Caledonia/20/1999	6	1.83	0.98	0.40	0.80
A/New Caledonia/20/1999	A/Fukushima/141/2006	6	3.17	0.75	0.31	0.61
A/Thessaloniki/24/2005	A/Solomon Islands/3/2006	6	3.50	1.52	0.62	1.24
A/New Caledonia/20/1999	A/Chile/8885/2002	6	0.67	0.52	0.21	0.42
A/Thessaloniki/24/2005	A/Hong Kong/2652/2006	6	2.83	0.41	0.17	0.33
A/Hong Kong/2652/2006	A/Fukushima/141/2006	6	0.83	0.41	0.17	0.33
A/Solomon Islands/3/2006	A/Thessaloniki/24/2005	6	3.83	0.41	0.17	0.33
A/Fukushima/141/2006	A/Hong Kong/2652/2006	6	1.17	0.41	0.17	0.33
A/Solomon Islands/3/2006	A/New Caledonia/20/1999	6	3.33	1.03	0.42	0.84
MEAN				0.75	0.28	0.56

Table 2.1: H1N1 one-way assays for which 6 or more results are available (data to 2008)

Strain	Strain	Assays	Log Distance	Std Dev	Std Error	95% conf +/-
A/Beijing/262/1995	A/New Caledonia/20/1999	14	1.39	0.49	0.13	0.26
A/Thessaloniki/24/2005	A/New Caledonia/20/1999	9	0.72	0.51	0.17	0.34
A/Johannesburg/82/1996	A/New Caledonia/20/1999	8	4.88	0.23	0.08	0.16
A/New Caledonia/20/1999	A/Egypt/96/2002	8	0.44	0.18	0.06	0.12
A/Johannesburg/82/1996	A/Beijing/262/1995	8	4.38	0.44	0.16	0.31
A/Hong Kong/2652/2006	A/Solomon Islands/3/2006	7	0.57	0.35	0.13	0.26
A/Fujian/156/2000	A/New Caledonia/20/1999	7	0.36	0.24	0.09	0.18
A/Hawaii/15/2001	A/New Caledonia/20/1999	7	2.57	0.45	0.17	0.34
A/Fujian/156/2000	A/Beijing/262/1995	7	1.86	0.24	0.09	0.18
A/Hong Kong/2652/2006	A/New Caledonia/20/1999	7	3.29	0.49	0.18	0.37
A/Solomon Islands/3/2006	A/New Caledonia/20/1999	6	2.67	0.82	0.33	0.67
A/Hawaii/15/2001	A/Beijing/262/1995	6	3.25	0.42	0.17	0.34
A/Chile/8885/2002	A/New Caledonia/20/1999	6	0.5	0.32	0.13	0.26
A/Thessaloniki/24/2005	A/Solomon Islands/3/2006	6	3.25	0.69	0.28	0.56
A/Hong Kong/2652/2006	A/Fukushima/141/2006	6	0.67	0.41	0.17	0.33
A/Hong Kong/2652/2006	A/Thessaloniki/24/2005	6	3.92	0.74	0.3	0.6
MEAN				0.44	0.17	0.33

Table 2.2: H1N1 two-way assays for which 6 or more results are available (data to 2008)

Strain	Antiserum	Assays	Log Distance	Std Dev	Std Error	95 % conf +/-
A/Wyoming/3/2003	A/Panama/2007/1999	21	1.95	0.74	0.16	0.32
A/Panama/2007/1999	A/Wyoming/3/2003	21	3.33	1.28	0.28	0.56
A/California/7/2004	A/Wyoming/3/2003	17	2.35	0.93	0.23	0.45
A/Wyoming/3/2003	A/California/7/2004	17	0.94	1.14	0.28	0.55
A/California/7/2004	A/Wisconsin/67/2005	17	1.88	0.99	0.24	0.48
A/Wisconsin/67/2005	A/California/7/2004	17	2.18	0.88	0.21	0.43
A/Panama/2007/1999	A/Moscow/10/1999	15	2.17	1.51	0.39	0.78
A/Moscow/10/1999	A/Panama/2007/1999	15	1.6	0.51	0.13	0.26
A/Moscow/10/1999	A/Sydney/5/1997	14	1	0.39	0.1	0.21
A/Panama/2007/1999	A/Sydney/5/1997	14	2	0.88	0.23	0.47
A/Sydney/5/1997	A/Moscow/10/1999	14	1.93	1.44	0.38	0.77
A/Sydney/5/1997	A/Panama/2007/1999	14	1.86	0.77	0.21	0.41
A/Wellington/1/2004	A/Wyoming/3/2003	13	1.54	0.88	0.24	0.49
A/Hiroshima/52/2005	A/California/7/2004	12	2.67	0.49	0.14	0.28
A/Wyoming/3/2003	A/Wellington/1/2004	12	1.33	1.15	0.33	0.67
A/Hong Kong/4443/2005	A/California/7/2004	12	3.58	0.9	0.26	0.52
A/New York/55/2004	A/California/7/2004	12	0.67	0.65	0.19	0.38
A/California/7/2004	A/New York/55/2004	12	1	1.04	0.3	0.6
A/Panama/2007/1999	A/Fujian/411/2002	12	4.67	0.78	0.22	0.45
A/Fujian/411/2002	A/Panama/2007/1999	12	4.08	1.62	0.47	0.94
A/Shantou/1219/2004	A/Wyoming/3/2003	11	4.18	0.75	0.23	0.45
A/Hong Kong/4443/2005	A/Wisconsin/67/2005	11	2.82	1.08	0.33	0.65
A/California/7/2004	A/Hong Kong/4443/2005	11	1.73	0.47	0.14	0.28
A/Wellington/1/2004	A/Panama/2007/1999	11	3.64	1.21	0.36	0.73
A/Shantou/1219/2004	A/Panama/2007/1999	11	5.18	0.87	0.26	0.53
A/Hiroshima/52/2005	A/Wisconsin/67/2005	11	0.45	0.69	0.21	0.41
A/Wisconsin/67/2005	A/Hong Kong/4443/2005	11	0.18	0.4	0.12	0.24
A/Wyoming/3/2003	A/Shantou/1219/2004	10	1.1	0.74	0.23	0.47
A/California/7/2004	A/Hiroshima/52/2005	10	3.3	0.82	0.26	0.52
A/Wellington/1/2004	A/California/7/2004	10	0.7	0.67	0.21	0.43
A/New York/55/2001	A/Panama/2007/1999	10	0.4	0.7	0.22	0.44
A/New York/55/2004	A/Wyoming/3/2003	10	2.5	0.85	0.27	0.54
A/Wyoming/3/2003	A/New York/55/2004	10	2.4	0.97	0.31	0.61
A/Wisconsin/67/2005	A/Hiroshima/52/2005	10	0.9	0.57	0.18	0.36
A/California/7/2004	A/Panama/2007/1999	10	5.7	1.06	0.33	0.67
A/Panama/2007/1999	A/Shantou/1219/2004	10	4.7	0.95	0.3	0.6
A/Panama/2007/1999	A/California/7/2004	10	5.5	1.18	0.37	0.75
A/Panama/2007/1999	A/Wellington/1/2004	10	3	0.82	0.26	0.52
A/California/7/2004	A/Wellington/1/2004	10	2	0.82	0.26	0.52
MEAN				0.89	0.25	0.51

Table 2.3: H3N2 one-way assays for which 10 or more results are available (data to 2008)

Strain	Strain	Assays	Log Distance	Std Dev	Std Error	95% conf +/-
A/Wyoming/3/2003	A/Panama/2007/1999	21.00	2.62	0.65	0.14	0.28
A/Wisconsin/67/2005	A/California/7/2004	17.00	1.82	0.58	0.14	0.28
A/California/7/2004	A/Wyoming/3/2003	17.00	1.32	0.53	0.13	0.26
A/Moscow/10/1999	A/Panama/2007/1999	15.00	1.18	0.80	0.21	0.42
A/Moscow/10/1999	A/Sydney/5/1997	14.00	0.93	0.92	0.25	0.49
A/Sydney/5/1997	A/Panama/2007/1999	14.00	1.25	0.51	0.14	0.27
A/New York/55/2004	A/California/7/2004	12.00	0.62	0.57	0.16	0.33
A/Wellington/1/2004	A/Wyoming/3/2003	12.00	1.42	0.60	0.17	0.34
A/Panama/2007/1999	A/Fujian/411/2002	12.00	3.71	0.66	0.19	0.38
A/Hong Kong/4443/2005	A/California/7/2004	11.00	1.77	0.26	0.08	0.16
A/Hong Kong/4443/2005	A/Wisconsin/67/2005	11.00	0.68	0.25	0.08	0.15
A/Shantou/1219/2004	A/Panama/2007/1999	10.00	4.00	0.85	0.27	0.54
A/California/7/2004	A/Wellington/1/2004	10.00	1.00	0.47	0.15	0.30
A/Hiroshima/52/2005	A/California/7/2004	10.00	3.00	0.53	0.17	0.33
A/Wellington/1/2004	A/Panama/2007/1999	10.00	3.25	0.54	0.17	0.34
A/Shantou/1219/2004	A/Wyoming/3/2003	10.00	1.60	0.52	0.16	0.33
A/Hiroshima/52/2005	A/Wisconsin/67/2005	10.00	0.55	0.37	0.12	0.23
A/New York/55/2004	A/Wyoming/3/2003	10.00	2.30	0.42	0.13	0.27
A/California/7/2004	A/Panama/2007/1999	10.00	5.65	1.33	0.42	0.84
MEAN				0.60	0.17	0.34

Table 2.4: H3N2 two-way assays for which 10 or more results are available (data to 2008)

Strain Pair	Measurement	Mean			Median		Standard Deviation		
		Calculated	Estimated	%	Calculated	%	Calculated	Estimated	%
A/Wyoming/3/2003 vs. A/Panama/2007/1999	One-Way	1.95	1.94	0.55	2	2.44	0.74	0.88	18.5
	Two-Way	2.62	2.62	0	2.5	4.55	0.65	0.68	5.13
A/Beijing/262/1995 vs. A/New Caledonia/20/1999	One-Way	1.29	1.23	4.67	1	22.2	0.61	0.89	45.3
	Two-Way	1.39	1.39	0	1.5	7.69	0.49	0.53	9.12

Table 2.5: Calculated and estimated values from selected strain pairs in the four data sets. Calculated values are derived directly from the dataset. Estimated values are estimated from the 'best fit' normal distribution according to the Normal Probability Plot. As mean and median values are equal for a normal distribution, the estimated mean value is also used as the estimated median.

2.2.4 Independence of the Observations

Surveillance laboratories often publish several HI assay tables in a season, utilising very similar sets of reference antisera. The question arises as to whether the titres between reference strains and antisera are determined *de novo* in each assay, or simply carried forward from one table to the next. Because of the dependency of results on red blood cell avidity, the assay will certainly be repeated anew when the supply of blood is changed (private discussions with Dr John McCauley and Dr Rodney Daniels, UK WHO Collaborating Centre for influenza surveillance,

November 2012). The supply is changed on a weekly basis, and inspection of the published tables in the database reveals them to be separated by at least this amount of time. Each assay recorded in the database can therefore be regarded as the result of a distinct laboratory experiment, or, more likely, the average of several such results obtained at the same time and under the same conditions. The viral samples used in the experiment are not, however, necessarily derived *de novo* from wild type viral specimens.

The antisera used in the HI tests collected in the database are predominantly 'reference antisera' prepared and distributed by the WHO Collaborating Centres, located in the US, UK, Japan, China and Australia. These are used in the reports of those centres, which form the very large majority of assays published since 2001, and are also used in many other assays published both since and before that date. Novel strains are typically isolated and cultured for a specific assay: however the Collaborating Centres provide reference strain cultures, which mirror the reference antisera. Assays of the reference strains against reference antisera are typically provided in each report of results, alongside assays of novel strains against the reference antisera, to act as a control. The reference strains and antisera are highly represented in the datasets of frequently-occurring measurements that we have been examining, particularly in the two-way datasets. While there is a significant degree of independence, since each Collaborating Centre develops its own reference material, and will develop several batches of a widely used strain over time, a degree of interdependence should be noted. For example, the Collaborating Centre in London published 5 H3N2 HI tables undertaken at different times between January and September 2006. The reference strain A/California/7/2004 was represented in each of these tests, and the assays of that strain against its antiserum from those 5 tables are recorded in the database as separate measurements. The note against the antiserum in each table (F31/05) indicates that the same source of antiserum was used in each case (that is, drawn from the same ferret).

In conclusion, the observations are independent in that they are separate observations in the laboratory. There is a degree of dependence in that antiserum from the same source is used in multiple experiments, particularly in those experiments for which there are large numbers of results. The conclusions that we draw from these results, and predictions that we make, are valid within the framework of the experimental system from which the data was drawn. We cannot guarantee their validity outside that framework.

2.2.5 Heteroscedasticity

Figure 2.2 shows plots of variance against $\log_2$ distance. Some outliers can be observed but there is no apparent dependence of variance on distance.

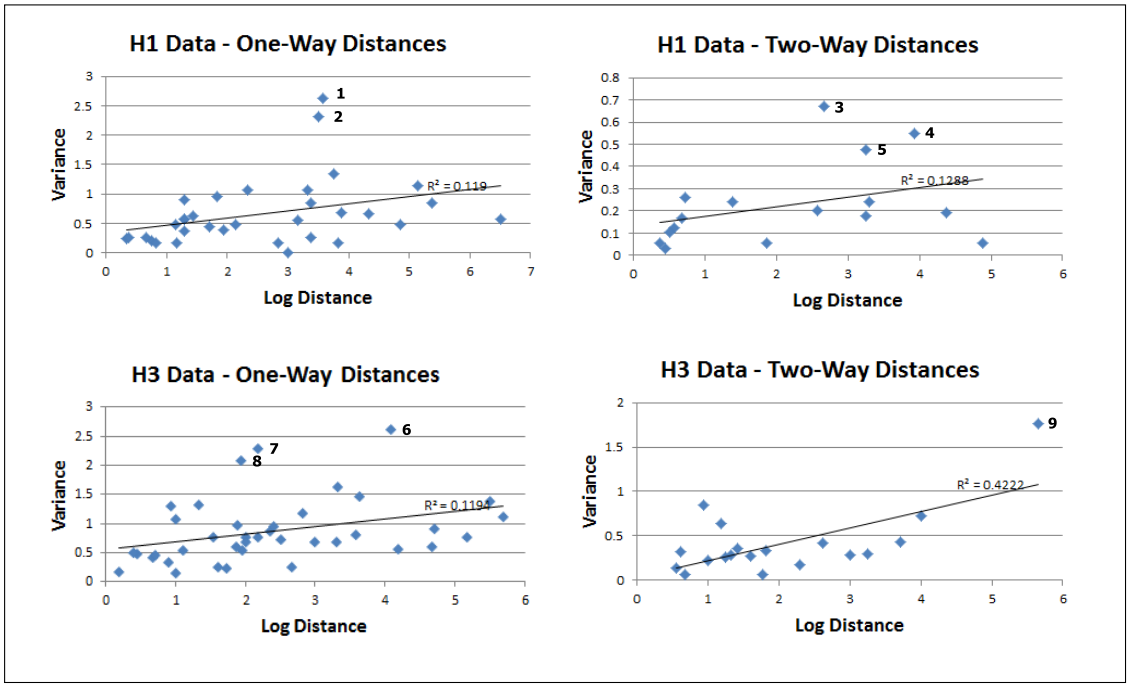


Figure 2.2: Plots of variance against log2 distance for the four data sets. Outliers are numbered.

The two-way measurements display generally lower variance than the one-way measurements, supporting the principle of normalisation underlying the Archetti-Horsfall calculation discussed in the opening section.

Nine outliers were identified in the plots. These are listed in Table 6. In most cases, the high variance was caused by one or in some cases two outliers in the underlying assay results, indicated by 'O' in the Notes column.

Outlier	Strain	Strain/Antiserum	Variance	Notes
1	A/New Caledonia/20/1999	A/Solomon Islands/20/2006	2.62	O
2	A/Thessaloniki/24/2005	A/Solomon Islands/20/2006	2.31	O/S
3	A/New Caledonia/20/1999	A/Solomon Islands/20/2006	0.67	O
4	A/Hong Kong//2652/2006	A/Thessaloniki/24/2005	0.55	O
5	A/Thessaloniki/24/2005	A/Solomon Islands/20/2006	0.48	O/S
6	A/Fujian/411/2002	A/Panama/2007/1999	2.62	P/S
7	A/Panama/2007/1999	A/Moscow/10/1999	2.28	O
8	A/Sydney/5/1997	A/Moscow/10/1999	2.07	O
9	A/California/7/2004	A/Panama/2007/1999	1.77	S

Table 2.6: Variance Outliers identified in the Variance/Log Distance Plots of Figure 2. In the Notes column, O denotes that variation is probably attributable to outliers in the results. S denotes variation that can probably be attributed to antiserum variability, and P to possible variations in experimental protocol (see text for discussion).

The presence of outliers in the underlying data suggests that a robust metric such as the median or M-estimate (Huber, 2005) will provide a better estimate of the log distance than would be provided by the mean. As an example, the data series for outlier 1 is 2.5, 4, 4, 4, 4.5, 4.5. The value of 2.5 appears to be an outlier. The mean is 3.92, which is below the value of the other data points. The M-estimate of the location using the R function huberM with default parameters is 4.09, which is more in keeping with the data set taken as a whole.

In outlier 9, the high variance appears to be caused by a change in the reference antiserum. Log distance values of 4, 4.5, 5, 5, 5, 5, 5.5 were obtained in 2005 using UK-supplied antiserum to A/Panama/2007/1999 reference F2/01. Values of 7.5, 7.5, 7.5 were obtained in 2006 using antiserum to the same strain with reference F7/06. All assays except for one assay in the 2005 series were performed by the same laboratory, and we are fortunate that the antiserum reference was quoted on all assays, including the single assay performed at another laboratory, as it is not usually included in the HI results. The likely cause of high variance in this case is noted as S in the table, for 'serum-related'.

The observation of the impact of the change in antiserum on this particular strain pair provides evidence that the consistency of the HI assay can be affected by variation in the immune responses of the animals from which antisera are drawn. On the other hand, the overall consistency of results obtained from different laboratories, for example the WHO Collaborating Centres in the UK and the US, which generally derive their own antisera independently, demonstrates that this is not the norm. In this particular case, the variation represented a relatively small (onefold or twofold) change in the result obtained for a strain/antiserum pair of two antigenically different strains. It may be that variation of this nature is more likely at high

antigenic distance, due to the correspondingly large differences in the antibodies likely to be raised against them.

Another interesting outlier is outlier 6 (strain A/Fujian/411/2002 vs. antiserum A/Panama/2007/1999), which has a wide spread of values (2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6) in which the lower results (2-4) appear to be derived from US/Australian-originated antiserum and the higher values (5-6) from UK-originated antiserum. However, in this case the divergence may well be related to differences in protocol. The mutations between these strains included changes in the RBS, and Fujian-like viruses proved difficult to work with, often providing low reactivity even in homologous titres as noted in the introductory section. The note against this outlier is 'P/S', indicating either a possible protocol difference, or a serum-related difference as for outlier 9.

Outliers 2 and 5 (both involving A/Thessaloniki/24/2005 and A/Solomon Islands/20/2006) exhibit a wide range of values (e.g. 1, 3, 3, 4, 5, 5 for outlier 2). These may be related to differences in antisera as the lower values were determined in 2007, and the higher in 2008, however in this case the reference antisera are not identified specifically and so one cannot be sure: another explanation would simply be that the value of 1 is an outlier. These lines are therefore marked 'O/S' in the table.

2.3 Analysis of sequence quality

An overall assessment of sequence quality was presented in previous work (Lees, 2009). In particular, I assessed the extent to which sequences of human strains may have been influenced by egg adaption in culture. In the database, I attempt to address this issue, and others impacting sequence quality, by deriving consensus sequences where multiple sequences are available for a strain, which is frequently the case for strains of interest such as vaccine strains and dominant wild-type strains. One important factor here is the standardisation of place names, such as South Australia/S. Australia. Again, this topic is addressed in detail in previous work.

For this study, I required full-length H1N1 and H3N2 HA1 and HA2 sequences for as many strains as possible, and in particular for prevalent wild-type strains. I wished to confirm high data quality for the prevalent strains, particularly in the early years. As described in detail in my previous work, viral sequences were extracted mainly from the NCBI's Influenza Virus Resource (Bao et al., 2008) and from the Influenza Research Database (Squires et al., 2012). These two resources provide curated, searchable databases of influenza sequences extracted

from GenBank (Benson et al., 2008). I found little variation in the human H1N1 and H3N2 sequences available from the two sources. Some sequences, particularly for early strains of interest, were not available from these sources and were extracted from published papers where available.

Another important resource, GISAID (http://platform.gisaid.org), provides strains to researchers under terms which do not allow for subsequent publication without explicit permission from the original contributing researchers. I elected not to download sequences in bulk from GISAID because of the possibility that we might wish to make the database publicly available at some point in the future. I did, however, consult GISAID for particular sequences of interest that were not available elsewhere, but did not obtain any hits.

2.3.1 Full-Length Sequences of Dominant Strains

The prevalent subtypes and circulating strains are reported periodically in the *Weekly Epidemiological Record*. The description is typically couched in terms of antigenic relatedness to reference strains. A typical example is shown in Figure 2.3. WHO vaccine strain recommendations are provided in the same journal.

(this figure is not included in the public version)

Figure 2.3: An extract from a report of influenza activity in the *Weekly Epidemiological Record* (No 9, 3 March 2006) describing the antigenic characteristics of influenza strains circulating in the previous season: in this case the Northern Hemisphere 2005/2006 season.

As will be noted in the cited characterisation of H3N2 strains, some seasons see a transition from one antigenic type to another, which will usually then become the prevalent antigenic type in the subsequent season (for brevity, I shall refer to the prevalent antigenic type, as characterised by a reference strain, as the prevalent or dominant strain). While the progression is clear (strains related to reference strain x are replaced by strains related to reference strain y),

the exact point at which a transition occurs is not always easy to pin down, and is known to occur at different times in different parts of the world (Russell et al., 2008). In Table 2.7 I show the prevalent circulating strains and recommended vaccine strains for each Northern Hemisphere season since 1968, as reported in the *Weekly Epidemiological Record*. I have necessarily made some simplifications in order to arrive at a single reference strain for each season, but the account accurately follows the progression of prevalent strains, and is in accordance with the account and approach taken by other researchers (Gupta, Earl, and Deem, 2006).

I conducted an analysis of sequences available for H3N2 predominant strains to 2005 (both partial and complete), identifying the source of each one and comparing multiple sequences where available. A total of 95 sequences were considered in this analysis, and multiple sequences were available for each strain. Generally, correspondence between sequences was good, with little variation between samples (typically 3 or fewer amino acid differences across the full-length protein), and I therefore have confidence in the consensus sequences. A number of sequences of early strains determined by Kostolanský et al. (2000) were found to vary significantly from those determined by other researchers (amino acid differences in 7-10 positions in HA1), and I therefore eliminated sequences from this source in this study. It should be noted that these sequences, which I feel to be misleading, have been utilised by other researchers. Russell et al. (2008) submitted sequences to GenBank which are identical to those of Kostolanský in the course of publishing their work, which I believe to be duplicates.

I found full-length sequences of A/England/42/1972 and A/Port Chalmers/1/1973, contributed by the NIAID Influenza Genome Sequencing Project, which were misnamed in GenBank as A/England/72 and A/Port Chalmers/73 (other sequences of these isolates covered HA1 only). I was able to confirm this error (private communication with Dr Yiming Bao, September 2012) and have requested the NCBI to correct the GenBank records. In the course of this work, I also identified 16 H3N2 sequences with an isolation date of 1968 which were categorised by GenBank and the Influenza Virus Resource as human wild-type strains, but turned out on examination to be laboratory-induced mutations of A/Hong Kong/1/1968 from a mouse adaption study. Again, I have requested the NCBI to correct the records of these strains.

The data quality issues described above emphasise the importance of careful curation if the correct results are to be obtained from a study of HA sequences, particularly those of early strains. I believe that this has received inadequate attention in a number of bioinformatics

studies that have used influenza HA sequences. It is worth noting that, thanks to the effort of mass sequencing initiatives, new full-length HA sequences (and, indeed, full viral genome sequences) for a number of early strains have been deposited with the NCBI during the period of this research.

Northern Hemisphere Season	H1N1		H3N2		Notes	Incidence in season W - widespread P - partial L - low or none		
	Prevalent	**Vaccine**	**Prevalent**	**Vaccine**		H1N1	H3N2	B
1968-69			A/Hong Kong/1968					
1969-70			A/Hong Kong/1968					
1970-71					mainly influenza B, isolated HK68			
1971-72			A/Hong Kong/1968		antigenically distinct strains reported, en72 emerging			
1972-73			A/England/42/1972					
1973-74			A/Port Chalmers/1/1973	A/England/42/1972			W	W
1974-75			A/Port Chalmers/1/1973	A/Port Chalmers/1/1973	A/Port Chalmers/1/1973 types emerging		W	L
1975-76			A/Victoria/3/1975	A/Port Chalmers/1/1973	also A/Scotland/840/1974		W	L
1976-77			A/Victoria/3/1975	A/Victoria/3/1975			P	L
1977-78	A/USSR/90/1977		A/Texas/1/1977	A/Victoria/3/1975		P	W	L
1978-79	A/USSR/90/1977		A/Texas/1/1977	A/Texas/1/1977	also A/Brazil/11/1978 (H1N1)	P	L	P
1979-80	A/Brazil/11/1978	A/USSR/90/1977	A/Bangkok/1/1979	A/Texas/1/1977	also A/Bangkok/2/1979	P	W	P
1980-81	A/Brazil/11/1978	A/Brazil/11/1978	A/Bangkok/1/1979	A/Bangkok/1/1979		P	W	P
1981-82	A/England/333/1980	A/Brazil/11/1978	A/Bangkok/1/1979	A/Bangkok/1/1979	little H3N2 activity	L	L	L
1982-83	A/England/333/1980	A/Brazil/11/1978	A/Philippines/2/1982	A/Bangkok/1/1979	A/Philippines/2/1982 (H3N2) emerging	L	P	L
1983-84	A/Chile/1/1983	A/Brazil/11/1978	A/Philippines/2/1982	A/Philippines/2/1982		P	L	P
1984-85	A/Chile/1/1983	A/Chile/1/1983	A/Philippines/2/1982	A/Philippines/2/1982		L	P	L
1985-86	A/Chile/1/1983	A/Chile/1/1983	A/Mississippi/1/1985	A/Philippines/2/1982	A/Christchurch/4/1985 (H3N2) also emerging	L	W	P
1986-87	A/Singapore/6/1986	A/Chile/1/1983	A/Mississippi/1/1985	A/Christchurch/4/1985		W	L	L
1987-88	A/Singapore/6/1986	A/Singapore/6/1986	A/Mississippi/1/1985	A/Leningrad/360/1986	A/Leningrad/360/1986 (H3N2) emerging	L	P	P
1988-89	A/Singapore/6/1986	A/Singapore/6/1986	A/Sichuan/2/1987	A/Sichuan/2/1987		W	W	P
1989-90	A/Singapore/6/1986	A/Singapore/6/1986	A/England/427/1988	A/Shanghai/11/1987	A/Beijing/353/1989 (H3N2) also emerging	L	W	P
1990-91	A/Singapore/6/1986	A/Singapore/6/1986	A/Beijing/353/1989	A/Guizhou/54/1989		P	P	P
1991-92	A/Singapore/6/1986	A/Singapore/6/1986	A/Beijing/353/1989	A/Beijing/353/1989		P	W	L
1992-93	A/Singapore/6/1986	A/Singapore/6/1986	A/Beijing/32/1992	A/Beijing/353/1989		L	P	P
1993-94	none	A/Singapore/6/1986	A/Beijing/32/1992	A/Beijing/32/1992		L	W	P
1994-95	A/Singapore/6/1986	A/Singapore/6/1986	A/Johannesburg/33/1994	A/Shangdong/9/1993		L	P	P
1995-96	A/Singapore/6/1986	A/Singapore/6/1986	A/Johannesburg/33/1994	A/Johannesburg/33/1994	A/Wuhan/359/1995 emerging	W	W	L

1996-97	A/Bayern/7/1995	A/Singapore/6/1986	A/Wuhan/359/1995	A/Wuhan/359/1995		L	W	P
1997-98	A/Beijing/262/1995	A/Bayern/7/1995	A/Sydney/5/1997	A/Wuhan/359/1995		P	W	L
1998-99	A/Bayern/7/1995	A/Beijing/262/1995	A/Sydney/5/1997	A/Sydney/5/1997		L	W	W
1999-2000	A/New Caledonia/20/1999	A/Beijing/262/1995	A/Moscow/10/1999	A/Sydney/5/1997	also A/Panama/2007/1999 (H3N2)	P	W	L
2000-01	A/New Caledonia/20/1999	A/New Caledonia/20/1999	A/Moscow/10/1999	A/Moscow/10/1999		W	L	P
2001-02	A/New Caledonia/20/1999	A/New Caledonia/20/1999	A/Moscow/10/1999	A/Moscow/10/1999		P	W	W
2002-03	A/New Caledonia/20/1999	A/New Caledonia/20/1999	A/Fujian/411/2002	A/Moscow/10/1999	A/Fujian/411/2002 (H3N2) emerging	L	P	W
2003-04	A/New Caledonia/20/1999	A/New Caledonia/20/1999	A/Fujian/411/2002	A/Moscow/10/1999		L	W	L
2004-05	A/New Caledonia/20/1999	A/New Caledonia/20/1999	A/California/7/2004	A/Fujian/411/2002	A/California/7/2004 (H3N2) emerging	L	W	P
2005-06	A/New Caledonia/20/1999	A/New Caledonia/20/1999	A/Wisconsin/67/2005	A/California/7/2004	A/Wisconsin/67/2005 (H3N2) emerging	L	W	P
2006-07	A/Solomon Islands/3/2006	A/New Caledonia/20/1999	A/Wisconsin/67/2005	A/Wisconsin/67/2005	A/Solomon Islands/3/2006 (H1N1) emerging	L	P	L
2007-08	A/Brisbane/59/2007	A/Solomon Islands/3/2006	A/Brisbane/10/2007	A/Wisconsin/67/2005	A/Brisbane/59/2007 (H1N1) emerging	W	P	L
2008-09	A/Brisbane/59/2007	A/Brisbane/59/2007	A/Brisbane/10/2007	A/Brisbane/10/2007		P	P	L
2009-10	A/California/7/2009	A/Brisbane/59/2007	A/Perth/16/2009	A/Brisbane/10/2007		W	L	P
2010-11	A/California/7/2009	A/California/7/2009	A/Perth/16/2009	A/Perth/16/2009		P	P	P

Table 2.7: Vaccine recommendations and circulating strains for each North Hemisphere season since 1968, as extracted from the *Weekly Epidemiological Record*.

3 Predicting Naturally Occurring Epitopes in Influenza A HA1

3.1 Aims and Overview

Given the amino acid sequences of the haemagglutinins of two influenza strains, one of which is a descendant of the other and known to be antigenically different, we would like to identify the amino acid substitutions in the descendant that are associated with antigenic escape. In this study, I examine the size, shape and composition of known epitopes, and develop a computational method for identifying HA1 substitutions between strains which match those properties. I apply this method to predict the likely epitopes in selected wild-type H1 and H3 strains, and derive a set of 'immunoactive[1]' H3 amino acid locations from this analysis, which is broadly consistent both with sets of locations derived via other techniques (discussed below in Section 3.4) and with the locations of known H3 epitopes. I develop a predictive model of antigenic distance based on this set of immunoactive locations, and demonstrate that it has improved specificity compared to models developed in previous studies.

The work in this chapter has been published: Lees WD, Moss DS, Shepherd AJ. (2011). Analysis of antigenically important residues in human influenza A virus in terms of B-cell epitopes. *Journal of Virology* 85, no. 17 (September): 8548–8555; although here I present an updated analysis of known epitopes using information added to the Protein Data Bank since publication.

3.2 B-cell Epitope Sizes in Influenza A

A recent structural analysis of a non-redundant set of 53 antibody-antigen complexes in the Protein Data Bank (PDB) (Berman et al., 2002) found that 75% of the epitopes consisted of between 15 and 25 amino acids and covered a contact surface area on the antigen of between 600 and 1,000Å^2 (Rubinstein et al., 2008). Previous mutation studies have demonstrated that a small number of the epitopic residues – typically between three and five (Novotny, 1991) – can contribute a majority of the binding energy, with the mutation of just a single key residue being sufficient in some cases to inhibit binding (Air, Laver, and Webster, 1990).

[1] I use the term 'immunoactive locations' to refer to those locations in the molecule at which an amino acid substitution is likely to lead to antigenic change.

I conducted an analysis of all crystal studies of influenza A / antibody complexes in the PDB as of January 2013. The contact surface areas on the antigen, as reported by the authors, ranged from 631 to 984Å^2. The methods used to determine the epitope were often unreported, and, given the considerable scope for variation in such approaches – both in the overall approach taken, and in the determination of cutoff values for significance (Sivalingam and Shepherd, 2012), for consistency, I used protein-protein interactions inferred by PDBsum (http://www.ebi.ac.uk/pdbsum/) to derive surface areas and epitopic residues in each structure. In PDBsum, hydrogen-bonded atomic contacts are inferred by appropriate distance and bond angle between donor/acceptor pairs, and non-bonded contacts (involving either a carbon or sulphur atom) by distance (Laskowski, 2009). The results (Table 3.1) are in agreement with the above-described more general analysis: the number of identified epitopic residues ranges from 14 to 22.

In order to understand the approximate dimension of an HA epitope, I determined the longest distance between constituent residues, as measured between C_α atoms. This ranged from 25Å to 58Å. The latter figure is high in comparison to measurements from other structures, the next highest distance being 45Å. It is associated with the complex of a broadly-binding antibody to an H3 strain in PDB structure 3ZTJ. The same antibody is shown in complex with an H1 strain in 3ZTN. This antibody's primary contacts are in the heavy chain variable region HCDR3 (no other HCDRs are involved). The conformation of HCDR3 in the H1 and H3 complexes is remarkably similar (Figure 3.1). Approximately two thirds of the interaction (in terms of buried surface area) is with the HA2 chain. The antibody light chain has an interaction with the HA2 of an adjacent monomer in the H3 complex, but not in the H1 complex. The difference may arise from an N-glycosylation site at HA1 38 present in H3 strains but not in H1 strains (Corti et al., 2011). Given the similarity of the heavy chain interaction between the two structures, it is likely to contribute the majority of the binding energy in both structures; however it is possible that the outlying residues are important in order to achieve the necessary orientation of the antibody in the presence of the oligosaccharide. From this description and the associated figures, it will be seen that this exceptionally large longest distance arises from the disjoint epitopic residues present in the H3 structure but not the H1 structure.

In the above analysis of longest distance, the distance between C_α atoms is measured directly, rather than across the protein surface. Arguably the latter might provide more realistic measurements in regions of the protein that are markedly ridged: however, examination of influenza A epitopes, which is confirmed in wider structural studies of B-cell epitopes (Laver et al., 1990; Kringelum et al., 2013), reveals them to be relatively flat.

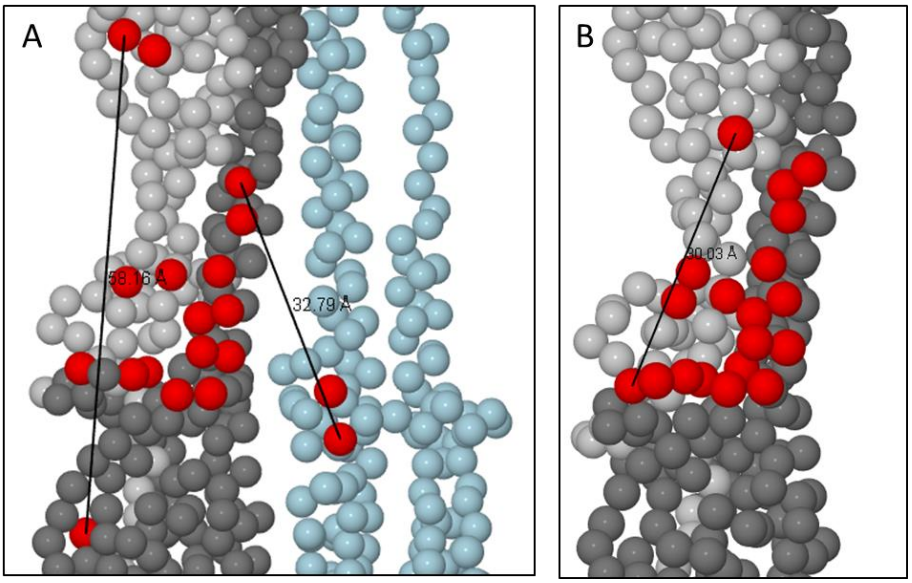


Figure 3.1: Partial structures of HA from PDB 3ZTJ (H3 subtype) (A) and 3ZTN (H1 subtype) (B): spheres indicate Cα atoms of amino acids constituting the HA protein. Red spheres indicate those identified by PDBsum as forming the epitope of antibody FI6V3 which binds to strains in both subtypes. HA1 is shown in light grey and HA2 in dark grey. In the H3 complex (A), an adjacent HA2 monomer is shown in light blue. The antibody binds across to this monomer, as well as employing outlying residues at both the membrane-distal (top of picture) and membrane-proximal sides of the core. These features are not observed in the H1 complex (B), although the structure of the core is quite similar

PDB Code	Subtype	Binding Region	Haemagglu-tination Inhibiting	Antibody Origin	Notes	HA Epitope (from PDBsum)				
						Reported S.A. (Å^2)	Longest Dist. (Å)	No. of Residues	HA1 locations	HA2 locations
2VIR	H3	Head	Yes	Mouse	Binds in sites A and B	631	25	17	129-137, 155-159, 190, 193, 194	
1KEN	H3	Head	Not Stated	Mouse	Binds across 2 HA monomers. Fusion inhibiting	884	29	18	128, 135-137, 156, 158, 165, 186, 187, 189, 190, 192-194, 222, 225-227	
1QFU	H3	Mid	Yes	Mouse		922	42	18	48-50, 59, 60, 62, 63, 74, 75, 78, 79, 90, 92, 94, 143, 271, 273, 274	
1EO8	H3	Mid	Yes	Mouse	Epitope substantially overlaps with 1QFU	787	28	15	50, 59, 60, 62, 63, 74, 75, 78, 79, 82, 90, 92, 94, 271, 273	
3GBN	H1	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	827	30	18	38, 40-42, 291, 292	19-21, 38, 41, 42, 45, 46, 49, 52, 53, 56
3GBM	H5	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	799	30	19	38, 40-42, 291-293, 318	19-21, 38, 41, 42, 45, 46, 49, 52, 53
3FKU	H5	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	723	29	14	32, 34, 292	18-21, 38, 41, 42, 45, 49, 52, 53
3SDY	H3	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	900	25	16	325	15, 16, 18, 19, 25, 30, 32-36, 38, 150, 153
3SM5	H1	Head	Not Stated	Human	Binds across RBS itself	775	30	20	134-137, 153, 155, 156, 158-160, 187, 189, 190, 192-194, 222, 225-227	
3ZTN	H1	Stalk	No	Human	Binds to H1 and H3 strains – same antibody as in 3ZTJ	911	30	19	28, 29, 289, 316	18-21,38, 39, 41-43, 45, 46, 49, 53, 56, 57
3ZTJ	H3	Stalk	No	Human	See above. H3 form binds across two monomers of HA	984	58	19	38, 277, 278, 318	18, 20, 21, 38, 39, 42, 43, 45, 46, 49, 53, 56, 57 (1st monomer); 7, 11 (2nd monomer)
3LZF	H1	Head	Not Stated	Human		946	31	18	125, 126, 128, 129, 157-163, 165-167,	

									169, 197, 246, 248	
4FQI	H5	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	782	32	19	38, 40, 41, 42, 291, 292, 293	18, 19, 20, 21, 36, 38, 41, 42, 45, 46, 48, 49
4FQV	H7	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	791	33	17	38, 291, 292	18, 19, 20, 21, 38, 41, 42, 45, 46, 48, 49, 52, 53, 56
4FQY	H3	Stalk	No	Human	Fusion inhibiting Binds across HA1, HA2	819	32	17	291, 318	18, 19, 20, 21, 37, 38, 41, 42, 45, 46, 48, 49, 52, 53, 56
4FP8	H3	Head (RBS)	Yes	Human	Broadly binding	626	22	18	98, 131, 133, 134, 135, 136, 137, 145, 153, 155, 156, 186, 189, 190, 192, 193, 194, 226	
4GMS	H3	Head (RBS)	Yes	Human	Broadly binding.	893	45	19	90, 131, 134, 135, 136, 137, 153, 155, 156, 157, 158, 159, 189, 193, 194, 196, 226	
			Mean			824	32.4	17.7		

Table 3.1: HA/antibody X-ray studies in the Protein Data Bank, showing the HA epitope surface area (i.e. the area buried on the surface of HA) and the longest distance between HA amino acids comprising the epitope. Surface areas and epitopic residues are calculated by PDBsum. Longest distances are derived from the protein structure, visualised in Jmol, and are measured between the Cα atoms of the corresponding residues.

3.3 Identified Clusters in H1 and H3 HA

I developed an algorithm to search for amino acid substitutions which 'cluster' within a specific distance of each other. The algorithm requires, as input, amino acid sequences of the two strains to be compared and a distance, hereafter referred to as the "cluster distance". The largest possible set of substitutions is found such that the C_α atoms of all substitutions in the set lie within the cluster distance from each other.

The first step in the calculation is to identify the substitutions between the two sequences. I then compute a distance matrix containing the distance between the C_α atoms of each pair of substitutions. Distances between C_α atoms are inferred from the X-ray structures of A/Aichi/2/68 (PDB code 1HGD) (Sauter et al., 1992) for H3N2 strains and A/Puerto Rico/8/34 (PDB code 1RU7) (Gershoni et al., 2007) for H1N1 strains. These structures were selected as being representative of the subtypes concerned, and, in contrast to many other available structures, because the HA is shown unbound from antibodies which might deform the gross structure. In the case of 1HGD, the HA is complexed with a synthetic construct in the RBS, but this is unlikely to cause more than small local deviations. Recently published crystal structures of 2004 and 2005 H3N2 HAs have revealed little change in backbone structure in H3 from 1968 to 2005 and validate the approach of using a single representative structure to locate and assess antigenic differences between strains (Lin et al., 2012).

The largest subset of substitutions meeting the required criterion – that the distance between any two substitutions in the set should be less than or equal to the cluster distance – is then discovered via a traversal of all possible permutations. If there are *n* substitutions, the algorithm will first check whether the set containing all *n* meets the criterion. If not, all permutations of *n*-1, *n*-2, … are successively checked, until, if necessary, all permutations containing at least three substitutions have been tested. If a conforming subset is identified, it is assigned to a cluster and the process is repeated with any remaining substitutions (i.e., discounting those that have already been assigned to a cluster). The process terminates when no further clusters can be identified.

Valid clusters are required to contain at least three substitutions, as 'clusters' of two substitutions are too common for those of significance to be distinguished. Both effective substitutions (i.e. those that become fixed in the viral population) and ineffective substitutions are considered for inclusion in clusters. In this work, they are distinguished following the approach and terminology of Shih et al. (2007): in summary, a substitution is classed as 'effective' provided that the newly dominant amino acid is present in at least 99% of samples

for a period of at least one year: an ‘ineffective’ substitution, by contrast, is one in which an amino acid becomes dominant at a location but never achieves 99% penetration in a year before ceasing to be dominant. It should be noted that a substitution classed as ineffective in this terminology, while not persisting in the long term, may nevertheless contribute to antigenic evolution in the shorter term, before being displaced by a more effective substitution.

The predominant circulating strains of influenza A H1N1 and H3N2 viruses in each year between 1972 and 2009 were identified from the annual and semi-annual influenza virus activity reports in the *Weekly Epidemiological Record* (http://www.who.int/wer/en/). The algorithm described in the above paragraphs was applied to each successive change in the predominant circulating strain in order to identify substitution clusters. Where multiple clusters were identified in substitutions between adjacent circulating strains, and where sequences of intervening strains were available, phylogenetic trees derived from sequence data using PhyML (Guindon and Gascuel, 2003) were used in conjunction with HI assay results compiled from data from journals and other sources to determine antigenic intermediates between the two epidemic strains. This allowed the evolutions of some multiple clusters to be separated. Identified clusters, using a cluster distance of 35Å, are shown in Figures 3.2 - 3.4. The cluster distance of 35Å was determined to be appropriate in this analysis (see Figure 3.8 and ensuing discussion), and is in good agreement with the crystallographically determined epitope dimensions described in the previous section and with the results of other studies (Rubinstein et al., 2008; Kringelum et al., 2013).

Interestingly, the H1 strains A/New Caledonia/20/1999 and A/Egypt/39/2005 are antigenically similar, although being separated by 6 years and 7 HA1 amino acid differences. There are a further 5 amino acid differences in HA1 between A/Egypt/39/2005 and A/Solomon Islands/3/2006, but the antigenic distance (shown in the diagram as a titre ratio rather than a log ratio) between these two strains is 7. In view of the low antigenic distance between A/New Caledonia/20/1999 and A/Egypt/39/2005, the six substitutions between these strains identified as a cluster in the diagram may represent successive dead ends on the path to eventual antigenic escape.

A question arises as to whether substitutions at ‘buried’ locations should be eliminated from consideration. Even if a location is identified as being buried in the reference structure, it is possible that a substitution could result in it becoming exposed: likewise, antibody binding could cause a local rearrangement resulting in the exposure of a buried location and its subsequent participation in the epitope. For these reasons, I did not discount buried locations when deriving clusters. Analysis of the structures 1HGD (H3) and 1RU7 (H1) using PISA

(Krissinel and Henrick, 2007) reveals a total of 21 buried locations in the H3 HA1 monomer and 27 in the H1 HA1 monomer. Two of these are identified as cluster members: H257Y in the H1N1 transition from A/USSR/90/1977 to A/Brazil/11/1978, and V202I in the H3N2 transition from A/Moscow/10/1999 to A/Hong Kong/1550/2002. The first is in a 5-location cluster and the second in an 11-location cluster: hence eliminating these two substitutions would not materially affect the results. Additional HA locations on the inward face of the monomer may not be accessible to antibodies in the trimeric structure, however few such locations are identified in my results (see Section 3.4 and Figure 3.7).

The predicted epitopes lie within closely defined regions of the monomer. All but three of the 21 H3N2 clusters and all but four of the 15 H1N1 clusters lie close to the RBS, with centroids positioned at 25Å or less from the membrane-distal end of the monomer. The exceptions are grouped in a region closer to the viral membrane, with centroids positioned between 40 and 55Å from the extreme membrane-distal end. I shall refer to these regions as the "RBS region" and the "mid region" in this work (Figure 3.5). Of the 76 amino acid locations participating in clustered substitutions between the H3N2 strains considered (both the series shown in Figure 3.3 and the series shown in Figure 3.4), 61 occurred within RBS clusters, and 18 occurred within mid region clusters, with three locations occurring in both clusters (Table 3.2).

Between successive H3N2 strains, I observed a number of substitutions in the mid region that did not meet the criteria for clusters developed above. I postulated that substitution clusters may develop in this region at a slower rate than in the RBS, in which case a snapshot between successive wild-type strains might not allow a sufficient window for a full cluster to develop. To take a coarser-grained view, I examined substitutions between representative strains of the 11 "antigenic clusters" identified by Smith et al. (2004), using the vaccine candidates identified in that work as being representative of each antigenic cluster.

The results are shown in Figure 3.4, and identify mid region clusters in 6 out of 10 cases. A possible seventh candidate can be seen in the transition from Wuhan/359/1995 to Sydney/5/1997, where the substitution at position 121 could potentially be a member of either predicted epitope but has been assigned to the RBS region due to the order in which the algorithm examines sets of substitutions.

Region	Amino Acid Identifiers
RBS	62, 75, 78, 80, 82, 83, 94, 96, 121, 122, 124, 126, 131, 133, 135, 137, 138, 139, 142, 143, 144, 145, 146, 155, 156, 157, 158, 159, 160, 163, 164, 182, 185, 186, 188, 189, 190, 192, 193, 194, 196, 197, 199, 201, 202, 207, 208, 213, 214, 216, 217, 219, 222, 223, 225, 226, 227, 233, 242, 244, 248
Mid	47, 50, 53, 54, 57, 62, 63, 82, 83, 172, 174, 260, 262, 275, 276, 278, 299, 308
Both	62, 82, 83

Table 3.2: Locations participating in H3 clusters identified in this study, classified by the region in which the cluster occurs

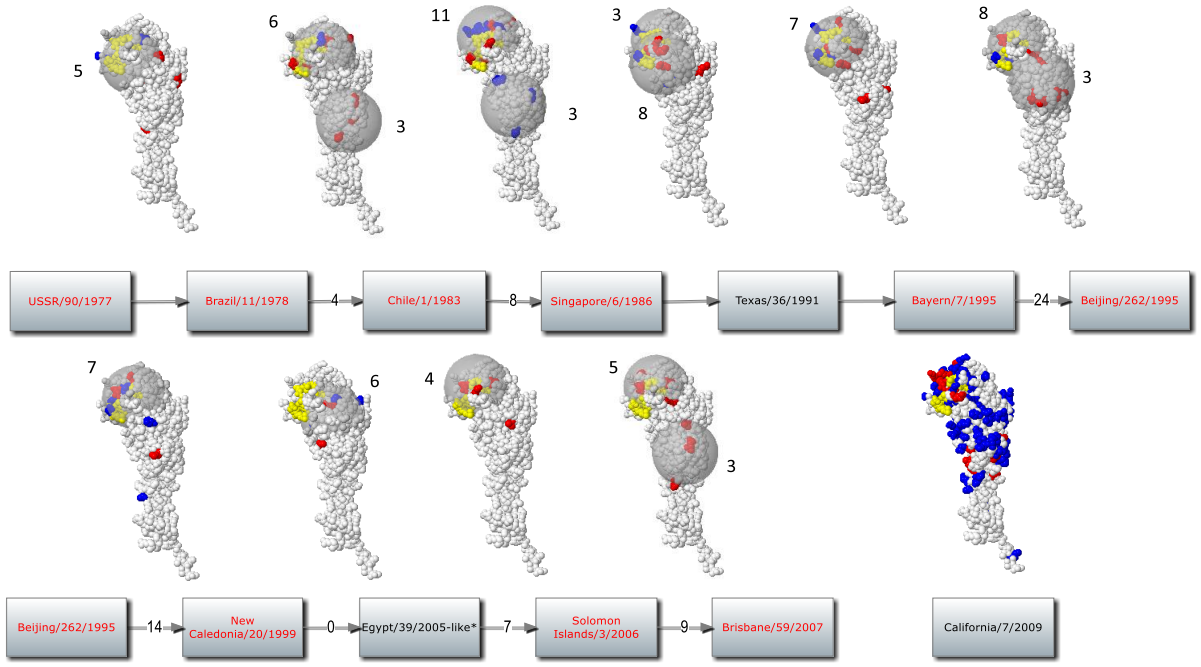


Figure 3.2: Clusters on the HA1 monomer calculated with a cluster distance of 35Å, for substitutions between selected H1N1 strains. Effective substitutions are shown in blue, and other substitutions in red. The Receptor Binding Site is shown in yellow. Clusters are designated by the light grey spheres. Predominating strains are listed in red text and intermediates in black text. The antigenic distance between the strain/serum on the left and the strain on the right is given in the arrow between the strain names where known from published results as the ratio between the homologous and heterologous HI titre. A value of 4 or more is generally regarded as immunologically significant (Schild et al., 1973). The number of substitutions in each cluster is given by the number near the cluster. The 2009 pandemic strain is shown for comparison, with amino acid differences compared to A/Brisbane/59/2007, although no cluster analysis has been performed given the large number of substitutions.

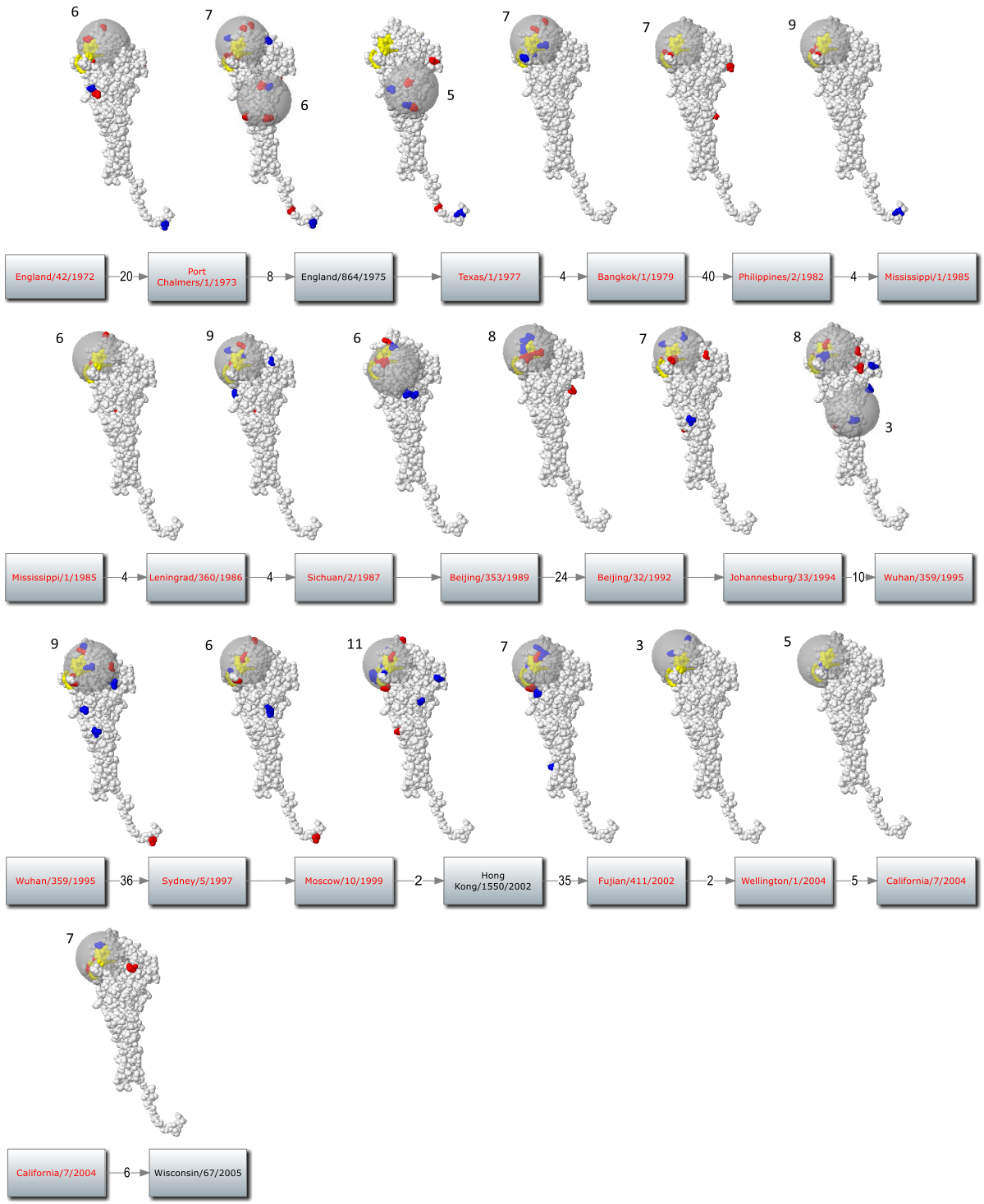


Figure 3.3: Clusters on the HA1 monomer calculated with a cluster distance of 35Å, for substitutions between selected H3N2 strains. Legend and colour coding as for Figure 3.2.

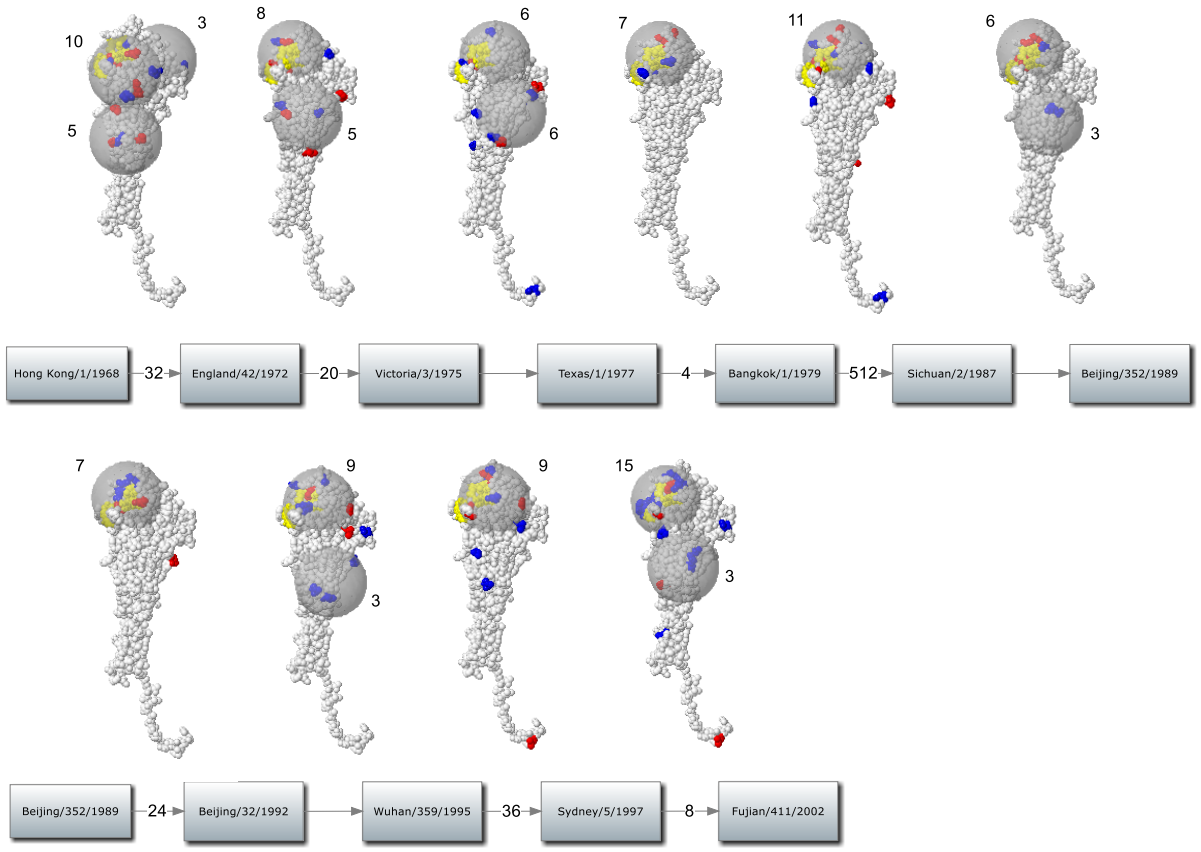


Figure 3.4: Clusters on the HA1 monomer calculated with a cluster distance of 35Å, for substitutions between H3N2 'antigenic clusters' (Smith et al., 2004). Legend and colour coding as for Figure 3.2.

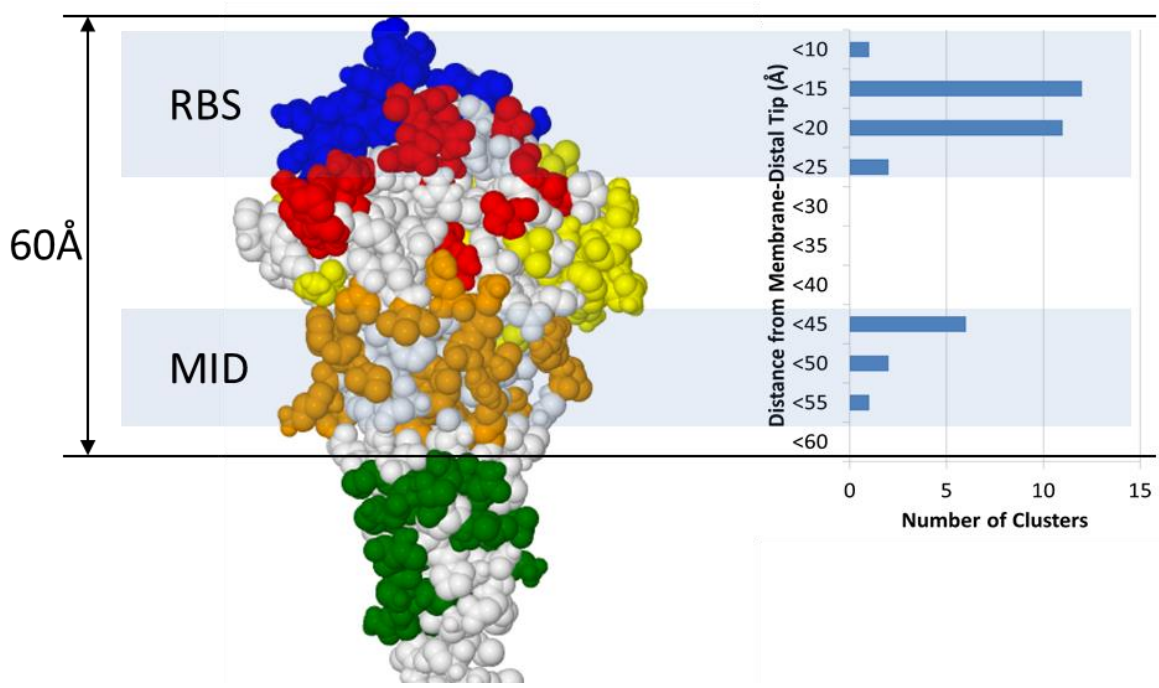


Figure 3.5: The centroids of identified clusters lie in two distinct regions of the H3 molecule, one close to the membrane-distal tip, and one at an axial distance of 40-55Å from the tip. In this study, these regions are termed the RBS and mid region respectively. The figure shows, for all H3N2 clusters identified in this study, the distribution of axial distance from the tip of the molecule to the cluster centroid, superimposed onto a single monomer of the global head of H3N2 (PDB structure 1HGD) for reference. Canonical H3 antigenic sites are shown in colour: A (red), B (blue), C (green), D (yellow), E (orange).

3.4 Comparison with H3 Canonical Sites

Table 3.3 lists the amino acid identifiers at which clustered substitutions were observed in the H3N2 strains considered, classified in terms of the canonical H3N2 antigenic sites. In previous work (Lees, 2009), we identified additional locations that should be considered members of these antigenic sites, in light of sequences not available at the time when the original list (Bush et al., 1999) was drawn up. Five locations are allocated to antigenic sites on the basis of this additional classification.

Of the 63 locations identified previously by Shih et al. (2007) as undergoing frequency switches, 57 are represented in the list of 76 substitutions identified in clustered substitutions. Twenty-three of the 25 locations identified by Yang (2000) as being under selective pressure (95% level, all models with listed locations) are included. Of the 45 locations identified previously by Smith et al. (2004) as being cluster differentiating, 41 are included, as are all 6 locations forming the decision tree constructed by Huang et al. (2011). These results give confidence that the regions identified by this technique are indeed those that are key to antigenic

differentiation, indicated both by evidence of selective pressure and by their importance in antigenic cluster differentiation (Figure 3.6).

Antigenic Site	**Amino Acid Identifiers**
A	**122*, 124*, 126, 131*, 133*, 135, 137***, 138, **142, 143*, 144*, 145*, 146*** (13/19)
B	**155*, 156*, 157**, **158*, 159, 160*, 163,** 164***, 186, 188*, 189*, 190*, 192, 193***, 194, **196*, 197*,** *199* (18/22)
C	47, **50*, 53*, 54*, 275, 276*, 278*, 299**, 308 (9/27)
D	96, **121, 172*,** 174*****, 182, 201***, 207***, 208, **213***, 214, 216, **217***, 219, ***222****, *223*, ***225****, **226**, **227,** *233*, **242**, **244***, **248** (22/41)
E	**57, 62*, 63**, **75*, 78**, 80, **82*, 83***, **94**, **260*, 262*** (11/22)
Unclassified	139, 185, **202*** (3/0)

Table 3.3: Locations participating in H3 clusters identified in this study, classified by canonical antigenic site (Bush et al., 1999). Locations in italics are additional locations not classified in that work but assigned subsequently to the cluster by virtue of their position (Lees, Moss, and Shepherd, 2010). Locations classified by Shih et al. (2007) as carrying effective frequency switches are in bold. Locations identified by Smith et al. (2004) as cluster differentiators are starred. Figures in brackets show the number of substitutions observed in the site in this study, and the total number of locations in the canonical H3 site.

The list is roughly half the size of the canonical list of varying locations in antigenic sites A to E. While this approach may miss some immunogenic locations (for example, in some cases escape may have occurred with substitutions of fewer than three residues), this significant reduction in number and the broad agreement with residues identified by other techniques suggest that not all of the canonical locations are immunogenically important.

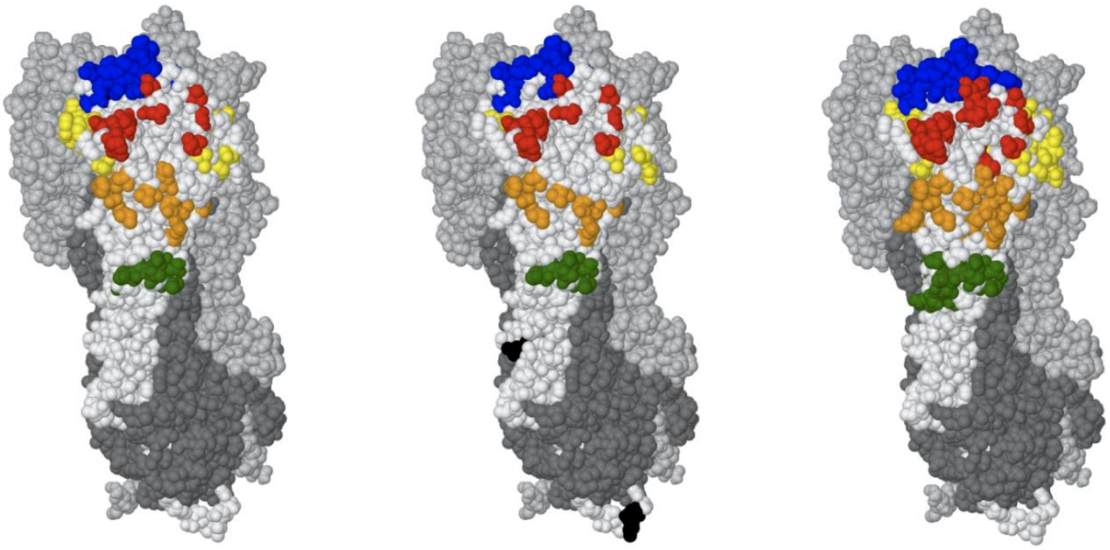
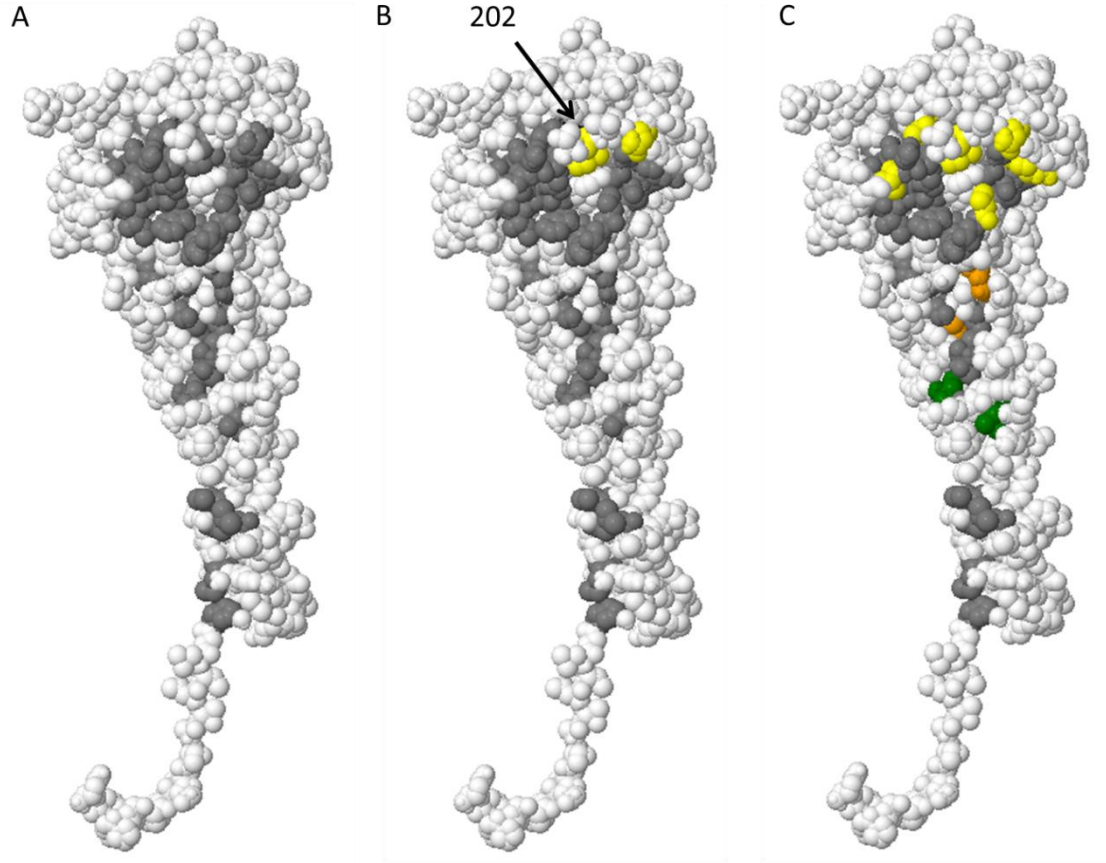


Figure 3.6: Comparison of Substitution Clusters with other sets of key residues. (A) Clustered substitutions observed in H3N2 strains considered in this study, superimposed onto one HA1 monomer (in white). (B) Locations undergoing effective frequency switches (Shih et al., 2007). (C) Canonical antigenic sites (Bush et al., 1999). Locations are coloured according to the antigenic site: A (red), B (blue), C (green), D (yellow), E (orange), or unclassified (black). The remaining two HA1 monomers in the complex are shown in light grey, and the three HA2 monomers are shown in dark grey. Some key residues are obscured in this view, and, indeed, obscured in the trimeric structure – see Figure 3.7 for a reverse view.

Some solvent-exposed residues in the HA1 monomer are not accessible to antibodies in the trimeric form, as they are occluded by HA2, or by other monomers. The exact extent of this occlusion cannot be determined with certainty, because the structure contains deep clefts and pockets: the extent to which antibodies can penetrate these is not known. To form an approximate view of the extent of the occluded region, I identified all HA1 residues whose C_α atom lay at a distance of 10Å or less from the C_α atom of a residue in HA2 or in another monomer, taking these to form the boundary of a region of potentially occluded residues. I then examined this region in spacefill view in Jmol, and removed from it all residues with apparent surface exposure in the trimeric form, leaving the set of occluded residues shown in Figure 3.7A. The set of 76 substitutions identified in this study includes three that lie in this occluded region (Figure 3.7B): all lie on the edge of the region closest to the membrane-distal tip. It is possible that residues in this location could contribute to antibody binding, either directly, via a minor conformational change, or indirectly (see Section 5.5 for a discussion of the role of buried residues in antibody escape). By contrast, the canonical set of 131 residues includes 11 which are distributed throughout the obscured region (Figure 3.7C). This is expected, as the authors considered solvent exposure but did not consider the blocking effect of the trimer as a whole. The low number of buried locations included in the set produced by my analysis provides some

confidence that the use of a directly measured cluster distance (as opposed to surface path) is



not introducing structurally questionable predictions.

Figure 3.7: Substitutions identified in this study lying on the inward face of the HA monomer, compared with those in canonical antigenic regions. (A) Residues shaded in grey are occluded in the trimeric structure. (B) Three residues identified in this study lie within the occluded region: residues 213 and 216 in antigenic site D, shown in yellow, and residue 202, which lies outside the canonical sites and is largely obscured in this view. (C) A total of 11 residues from the canonical set lie within the occluded region (Legend and colour coding as for Figure 3.2).

3.5 Cluster distance versus epitope size

To obtain an understanding of the sensitivity of this approach to the cluster distance, I examined the proportions of total and effective substitutions lying within calculated clusters for cluster distances of between 20Å and 60Å. The results are presented in Figure 3.8. With a cluster distance of 35Å, approximately 80% of the substitutions between the strain transitions illustrated in Figures 3.2-3.4 were contained within clusters. An increase of the cluster distance from 35Å to 60Å does not increase the percentage significantly, indicating that the remaining substitutions are sufficiently distant from clusters that they are unlikely to form part of a

common region. I therefore consider a distance of 35Å to be a reasonable trade-off between inclusion and specificity: however, in the spirit of the method, this is not a highly optimised cut-off. Interestingly, the same pattern can be seen both for total substitutions and for effective substitutions.

To understand the significance of these results, I compared them with simulated results obtained by distributing substitutions at random on the H3 monomer, considering all possible locations in HA1. The H3 series contained 19 strain transitions with between 3 and 16 substitutions in each transition. I used bootstrap tests to examine the significance of the clusters of substitutions obtained compared to those expected from a random distribution of substitutions across the HA1 monomer. In these tests, 1,000 batches of 19 simulated strain transitions were created, with the number of substitutions reflecting the distribution of the number in the H3 series and with the substitutions positioned randomly across the monomer. I observed a significant increase in the number of substitutions in a cluster at all diameters between 25Å and 60Å ($p < 0.01$) compared to a random distribution of substitutions. Even if the randomly assigned substitutions were chosen just from the 131 canonical residues in the five antigenic sites, this level of significance was found for H3 cluster distances of between 25Å and 45Å (the mean number of substitutions in an H3 cluster at 35Å was 6.8 in my results, compared to 5.3 in the random simulation with substitutions selected from the 131 locations). The clusters obtained from the analysis contain a significantly higher number of residues than would be expected from a random distribution, and are therefore unlikely to be just an artefact of sample size.

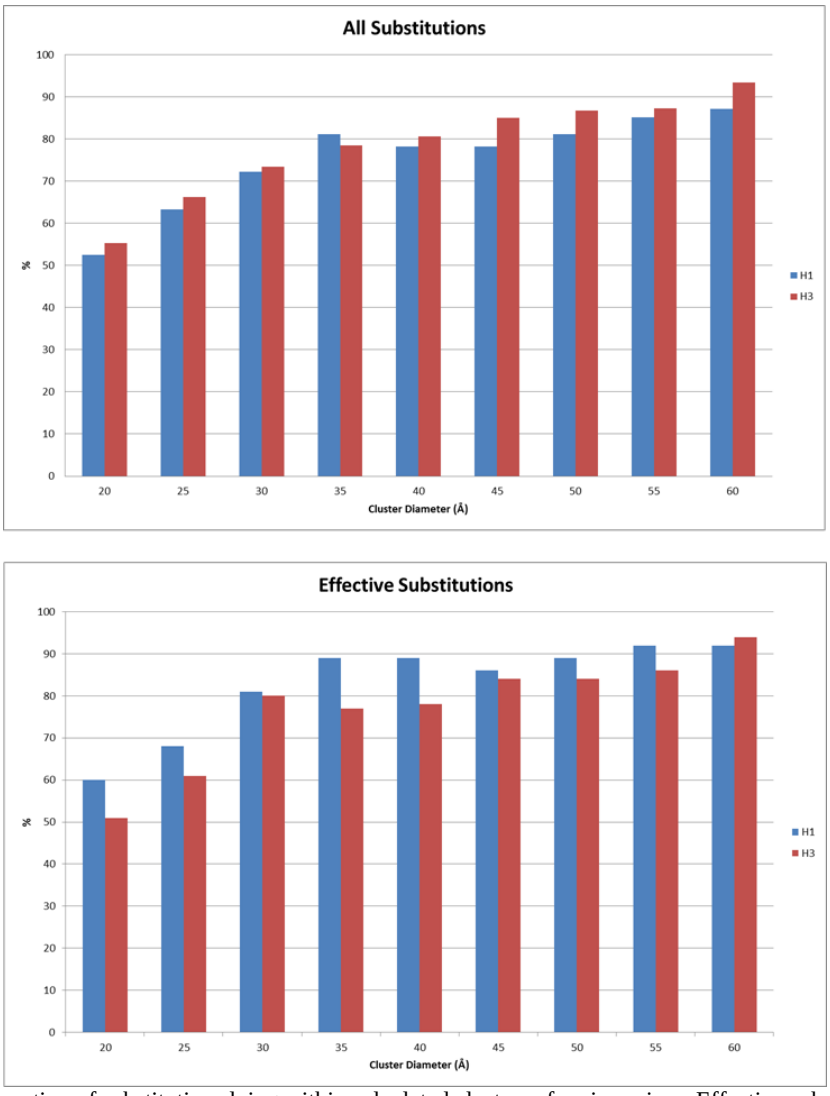


Figure 3.8: Proportion of substitutions lying within calculated clusters of various sizes. Effective substitutions follow the definition described previously by Shih et al. (2007).

3.6 Comparison of Cluster Locations with Locations of Known Epitopes

Okada et al. (2010) isolated and cloned 98 antibodies to wild-type H3 HA from a single volunteer born in 1960. These antibodies were found to divide into three sets: one set that bound to strains isolated between 1968 and 1973, a second set that bound to strains isolated between 1977 and 1993, and a third set that bound to strains isolated between 1993 and 2003. This experimental study used a chimeric approach in which various regions of the HA1 protein of a strain of interest are replaced by equivalent regions from a known antigenic variant. By this means, it is possible to test if the epitope of an antibody known to bind to the strain of interest contains critical residues in the replaced regions. The replacement regions varied in length from 5 to 32 amino acids. The technique is not able to identify the specific location of critical residues, but can isolate them to the substitutions within any replacement region found to elicit antigenic escape.

I compared the residue locations obtained by Okada et al. with 35Å clusters calculated from a predominant strain transition at or just after the end of each of the three identified periods of antibody binding, reasoning that substitutions in this transition would have led to escape.

Okada et al. isolated 11 antibodies binding to viral strains isolated between 1968 and 1973: I compared their binding locations with clusters calculated for the transition between A/England/42/1972 and A/Port Chalmers/1/1973. Nine of the antibodies bound across antigenic sites B and D in a location that is consistent with that identified in this analysis. The remaining two bound in the RBS region, inside the cluster identified by my results (Figure 3.9A). In the period of 1977 to 1993, most of the isolated antibodies bound in a region close to a mid region cluster that was calculated for the transition between A/Beijing/32/1992 and A/Wuhan/359/1995. The remainder bound in site E and site B. The set of locations identified in site E lie within the calculated mid region cluster, while those identified in site B lie within the calculated RBS cluster. My analysis predicts an epitope in the RBS region in this period: although such an antibody was not isolated from the experimental subject, it is possible that such antibodies were present in the wider population (Figure 3.9B).

In the period of 1993 to 2004, most antibodies isolated in the study showed binding activity in site B. Two of the substitutions in the cluster that were calculated for the transition from A/Fujian/411/2002 to A/Wellington/1/2004 are included in the set of residues from the study. The remaining antibodies from the study bound in the mid region; no substitutions in this region were observed for this strain transition. The experimental study also identified some binding activity in site A: the identified locations lie within the calculated cluster (Figure 3.9C).

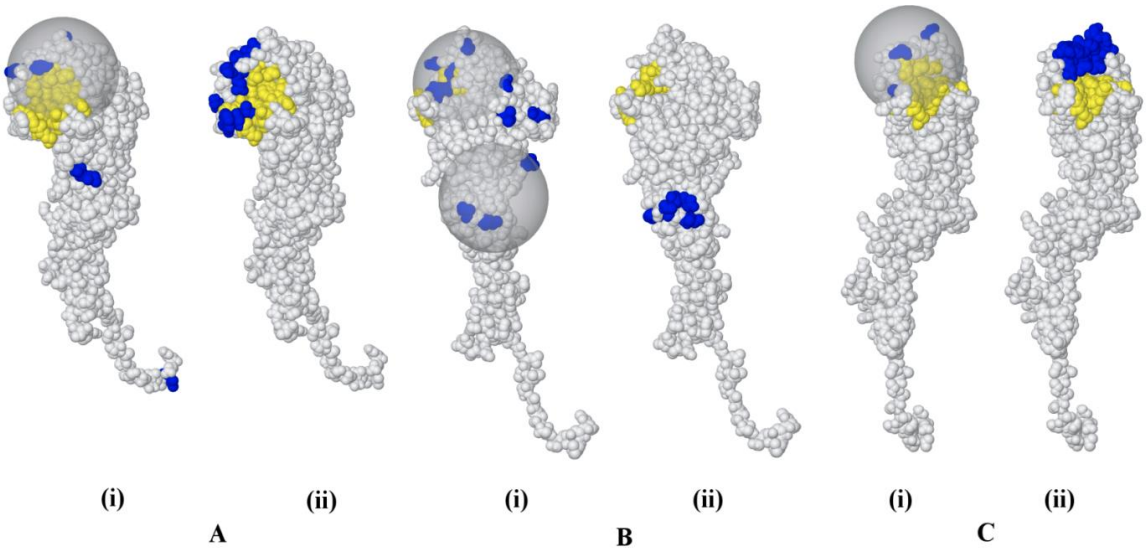


Figure 3.9: Comparison of clusters obtained in this analysis with data from an experimental study of antibodies isolated from a single individual (Okada et al., 2010). In each case, panel (i) shows the clusters that were obtained from a transition between the predominant wild-type strain and panel (ii) shows those residues identified in an experimentally defined chimeric approach, at least some of which disrupt the binding of an antibody that became ineffective after that transition. (A) Transition of A/England/42/1972 to A/Port Chalmers/1/1973; (B) transition of A/Beijing/32/1992 to A/Wuhan/359/1995 (B(ii) shows the mid region only); (C) transition of A/Fujian/411/2002 to A/Wellington/1/2003. Substitutions are in blue, and the receptor binding site is in yellow.

The antibody complexes in the PDB, summarized earlier in Table 3.1, include nine H3 structures, with epitopes spanning the head, mid and stalk regions. Some interesting comparisons can be drawn between these structures, my results, and the results reported by Okada et al. Turning first to the four H3 structures displaying head-region epitopes, the structure 2VIR binds in antigenic sites A and B, and in terms of location, is typical of many of the clusters that I have identified in H1 and H3 strain transitions and the RBS binding antibodies identified previously by Okada et al. The structure 1KEN binds across antigenic sites A, B, and D in one HA monomer and sites A and B in another. The binding location is again typical, but this technique and that reported previously by Okada et al. do not specifically address cross-monomer binding. The antibodies in structures 4FP8 and 4GMS bind across the RBS, with contacts in antigenic sites A, B and D. These antibodies are broadly binding to H3 strains and to other subtypes. Interestingly, though, their epitopes include many locations identified in Table 3.3: 14 out of the 18 residues in the 4FP8 epitope are in this list, as are 12 out of the 19 residues in the 4GMS epitope. Other residues in these two epitopes are strongly conserved, and may form the critical contacts allowing for their broad spectrum action.

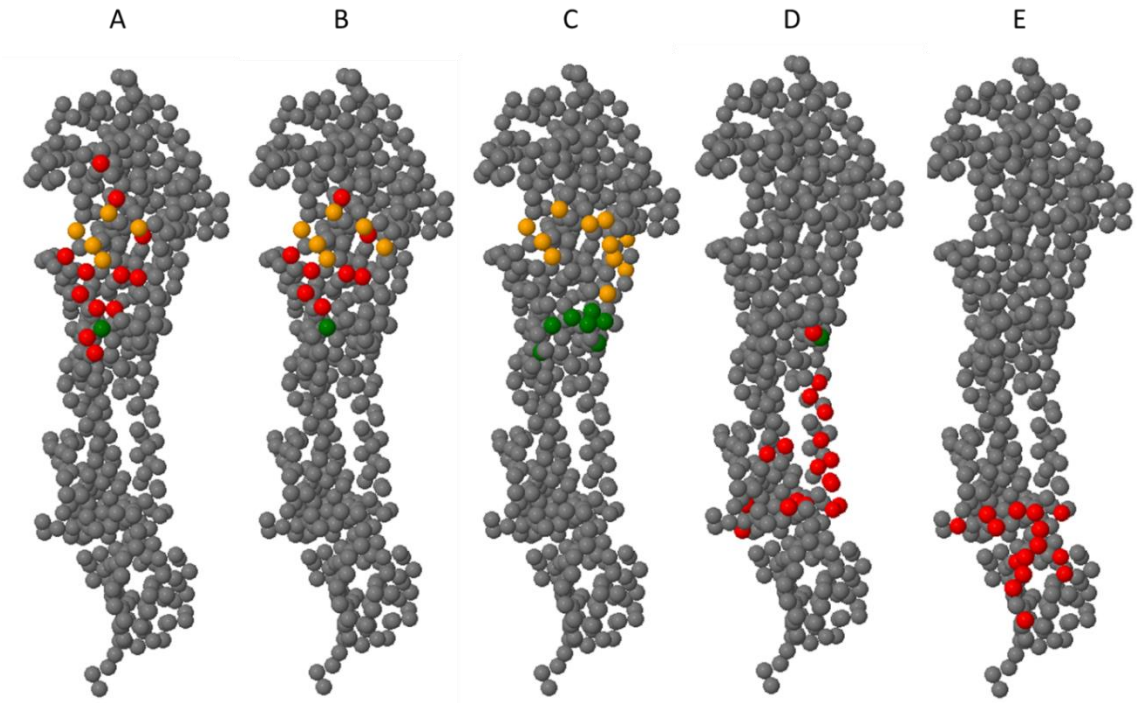


Figure 3.10: Comparison of H3 epitopes deduced from structures reported under PDB codes 1QFU (A), 1EO8 (B), 3ZTJ (D), and 3SDY (E) compared to the locations identified in this analysis in canonical antigenic sites C and E (C) (site E in yellow, site C in green). As in Figure 3.1, spheres indicate Cα atoms of amino acids constituting the HA protein. Yellow and green spheres in (A) and (B) show those locations in the epitope that are also found in (C). Red spheres in (A),(B),(D) and (E) indicate those identified by PDBsum as forming the epitope of the respective antibody that are not identified in (C). All epitopes highlighted on structure 1EO8.

Figure 3.10 compares two mid region and two stalk-region epitopes identified from H3 structures in the PDB with the antigenic locations identified from this analysis in canonical sites C and E. There is good correspondence between the mid region epitopes and the identified locations, with both epitopes encompassing the two antigenic sites. The same 6 HA1 locations in each epitope are included in the set of immunoactive locations identified by my technique: location 50 in antigenic site C, and 62, 63, 75, 78, 94 in antigenic site E. The five locations identified in the mid region in the Okada study (Figure 3.9, Bii , HA1 locations 50,53,54,276,278), at least some of which are escape locations for antibodies isolated in that study, are all locations identified by my technique.

The two stalk-region epitopes identified from H3 structures shown in the figure bind close to the viral membrane, although in 3ZTJ, the outlier at HA1 278 is included in the antigenic locations in site C identified by cluster analysis. Given the preponderance of HA2 locations in these two epitopes, it is not surprising that they do not coincide with locations identified by cluster analysis, as that analysis was confined to HA1. The H1 and H5 stalk-binding epitopes identified in crystallographic studies lie in equivalent locations between those of the two H3 epitopes, and again have a preponderance of HA2 residues. A further, final, H3 stalk-binding antibody,

identified in PDB structure 4FKY, binds in a similar region to that of 3ZTJ (see Section 5.6 for further discussion of stalk binding antibodies to H3).

An interesting question arises as to whether the orientation of the antibody isolated by Okada et al. whose antigenic escape region is depicted in Figure 3.9B(ii) is closer to that of the two H3 mid region epitopes 1EO8 and 1QFU, or to that of the stalk-binding epitope of 3ZTJ. The Okada study did not assess HA2 locations for antigenic escape, but the antibody was found to be neutralising and not haemagglutination inhibiting. These points suggest an orientation for this antibody closer to the viral membrane than that of the mid region antibodies, firstly as the latter are still sufficiently close to the RBS to inhibit haemagglutination, and also because, in the absence of steric interference with the RBS, the neutralisation mechanism is likely to be inhibition of the fusogenic transition, which requires binding to the HA2 chain in the stalk (Yewdell, 2011) . From Table 3.1 it can be seen that, for the set of known structures, antibodies binding exclusively to HA1 are haemagglutination inhibiting, while those binding also to HA2 are not, presumably because their increased distance from the RBS prevents steric interference with binding.

To obtain a wider comparison between the antigenically active regions identified in this study and those obtained by other researchers, I conducted a search of the Immune Epitope Database (http://www.immuneepitope.org/) (Vita et al., 2010) for conformational epitopes on human H1 and H3 HA1. Antibodies raised against synthetic peptides were excluded. I obtained references for a total of 16 additional studies (Table 3.4). Epitopes in the region of the RBS were found in 15 of those studies, and the locations identified were in good agreement with the locations identified in this work. Mid region epitopes were found in three of those studies. Two of these mid region epitopes lie across HA1 and HA2; the third is confined to HA1 in the same region as that identified by my analysis and by Okada et al. and shown in Fig. 6C. One study identified an epitope that is distinct from those identified in cluster analysis; interestingly, this epitope is from an antibody isolated from a human volunteer which was found to bind to a relatively conserved region of H3 HA at positions 173 to 181 (Kubota-Koketsu et al., 2009).

Subtype	Host	Strain	RBS	Mid	Other	Ref
H1	mouse	1977-1986, various strains	Y			Yamada et al. (1991)
H3	mouse	X31	Y	Y		Smith et al. (1991)
H1	mouse	A/PR/8/34	Y			Stark and Caton (1991)
H1	mouse	A/PR/8/34 and escape variants	Y			Meek et al. (1989)
H3	mouse	1968-1977, 5 strains	Y		Y	Underwood (1984)
H1, H2, H3	mouse	Various		Y		Okuno et al. (1993)
H3	mouse	X31	Y			Laeeq et al. (1997)
H1	mouse	Beijing 262/95, New Caledonia 20/99	Y			Morrissey and Downard (2006)
H3	mouse	Panama 2007/99,Shangdong 9/93,Kiev 301/94,	Y	Y		Morrissey et al. (2007)
H3	mouse	A/Kamata/14/91(H3N2)	Y			Nakajima et al. (2007)
H1	human	Pandemic 1918	Y			Yu et al. (2008)
H3	mouse	Panama 2007/99,Shangdong 9/93,Kiev 301/94,	Y			Morrissey and Downard (2008)
H1,H3	human	Various	Y		Y	Kubota-Koketsu et al. (2009)
H1	mouse	A/New Caledonia/20/99	Y			Schwahn and Downard (2009)
H1	human	Pandemic 1918	Y			Krause et al. (2010)
H1	human	Pandemic 1918 and 2009	Y			Xu et al. (2010)

Table 3.4: H1 and H3 HA1 B-cell epitopes identified from references extracted from the Immune Epitope Database (search conducted on 22nd May 2011). For each one, the epitope's location, derived from locational information provided in the database, is listed. Typically the database identifies a small number of locations experimentally determined to lie within the epitope.

3.7 Predictive Models of Antigenic Escape Based on Identified Cluster Participants

A viral strain of influenza A is considered to be antigenically different to antiserum raised from a different strain if the antigenic distance rises above a certain threshold, typically 2 (corresponding to a titre ratio in excess of 4). The appearance of strains that are antigenically different to the vaccine antiserum is an indication that the vaccine strain may require updating (Schild et al., 1973; Smith et al., 2004; Carrat and Flahault, 2007).

Researchers have previously noted an approximately linear relationship between the antigenic distance separating two H3N2 strains and various counts of their HA1 amino acid differences (Lee and Chen, 2004; Liao et al., 2008). They have utilised this relationship to build predictive models of antigenic difference or similarity. Such models are potentially useful for vaccine selection, and, being binary classifiers, are simple to evaluate. Lee and Chen (2004) obtained best performance by counting differences at the canonical set of 131 locations lying close to known antigenic sites and known to exhibit variation (Bush et al., 1999). Liao et al. (2008), in their most successful predictive model, improved overall performance by introducing two refinements. Firstly, amino acid differences were ignored unless the type of amino acid (classified as polar, nonpolar, positively charged, or negatively charged) changed. Secondly, multiple regression was used to select immunoactive locations from the full set of 131 locations used by Lee and Chen, and only changes at these locations were considered. Using the training and validation data assembled by Lee and Chen, and introducing these refinements, they achieved an agreement rate in the validation data set of 92% compared to Lee and Chen's 83%, yielding a sensitivity of 84% and specificity of 94% (Lee and Chen did not provide sensitivity or specificity scores).

In previous work (Lees, Moss, and Shepherd, 2010), we widened the canonical set of 131 locations to include a further 110 locations neighbouring the known antigenic sites and exhibiting variation in later years than those considered by Bush et al. in 1999. To incorporate locational information in our model, we developed a novel approach in that work, in which these locations were divided in to bands according to their distance from the membrane-distal tip of the protein, with the count of the number of substitutions in each band being used as the explanatory variables of a linear model. We also incorporated terms in the model to account for changes in N-glycosylation between the two strains. With our models, we were able to exceed the sensitivity of the models of Liao et al. but obtained generally poorer specificity. On the training and validation sets used in that study, which varied from those of Liao et al. and in particular strictly separated the strains used in training and validation sets, our best model obtained a sensitivity of 97% and specificity of 57%, compared to the approach of Liao et al.

which achieved sensitivity 83%, specificity 73%. Our models have an advantage in terms of generality, in that they do not rely on multiple regression to narrow the locations of interest to those evidenced in the training set, but the low specificity suggests that substitutions that do not relate to antigenic change are being included. In our models, we did not observe an improvement through discounting substitutions that did not change amino acid type.

3.8 Predictive Models Based on Cluster Participation

I reasoned that those locations identified in this study as participating in antigenic clusters might form a useful and relatively generalised set of locations on which to base a predictive model. To provide a representation of the spatial distribution of the amino acid differences, analogous to the bands in previous work, I superimposed a three-dimensional grid onto the HA1 molecule, using the reference coordinates from the X-ray structure of A/X-31 (PDB code 1HGD) (Sauter et al., 1992). Each amino acid was assigned to the cell of the grid in which its C_α atom is found. Changed amino acids were counted in each cell; for example, if, between two strains, there were three amino acid substitutions within a cell, the "difference" count for that cell would be 3. I tested models with a range of cell sizes, as described later.

The counts obtained were used as explanatory variables in a linear model of the form

$$D_{ij}^{c} = \sum_{n} x_n c_n^{ij} + k$$

where D_{ij}^{c} is the antigenic distance between the strains *i* and *j* as calculated by the model; the sum is over all cells in the grid, c_n^{ij} is the count of the number of amino acid differences between the strains in cell *n*, and the variables x_n and *k* are parameters to be determined by minimizing the least-squares residual given by

$$S = \sum (D_{ij}^{o} - D_{ij}^{c})^2$$

calculated over a training set, where D_{ij}^{o} is the observed distance (from HI assay data) between the strains as reported in the training set. The R module lm() (http://www.R-project.org) was used to perform calculations.

The ability of the model to correctly predict antigenic similarity or difference (i.e., antigenic distance <=2 or >2), was tested on a validation data set. In order to provide a benchmark comparison of performance, I used the training and testing pairwise data set reported by Liao et al. (2008). The training set consists of 181 antigenic distances measured between 45 viruses

isolated between 1971 and 2002, and the validation set consists of 96 measurements between 19 viruses isolated between 1999 and 2003. Two commonly used reference strains (A/Panama/2007/99 and A/Fujian/411/2002) are included in both datasets. While the inclusion of measurements against these strains in both training and validation sets is not ideal, this dataset was used in order to allow comparison of results with previous studies. I present results based both on the distances used in that study and on an extended set, in which the strain pairs considered are maintained but the quality of the distance measurements is improved by including additional titres from published studies not considered by Liao et al.

Being conscious that the selection of a particular cell size could lead to over-fitting, I reviewed the impact of cell size on predictive properties. Results are presented for grids with cubical cells with edge lengths varying between 2Å and 22Å. At 8Å, 47 cells of the grid were occupied, giving an average of 1.6 residues per cell. At 22Å, 10 cells were occupied, giving an average of 7.6 residues per cell. The predictive quality of the models was assessed by calculating the Matthews correlation coefficient (MCC) (Matthews, 1975). Figure 3.11(A) shows the predictive performance of the model at various cell sizes. I was interested to know the extent to which substitutions in the mid region contributed to predictive performance. I therefore constructed a second model, in which only the 61 identified residues in the RBS region were considered. Results for this model are presented in Figure 3.11(B).

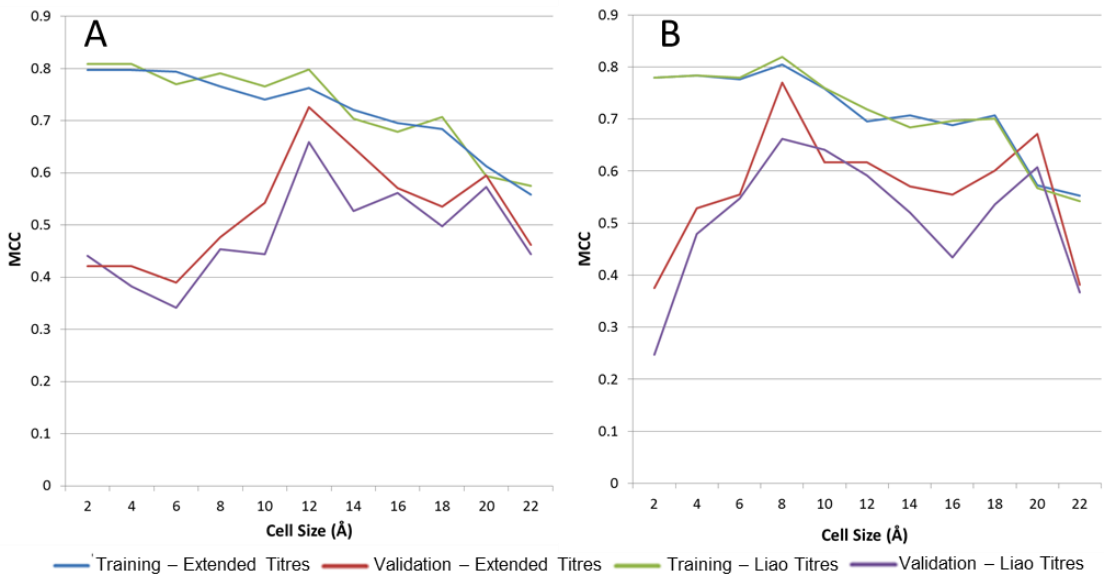


Figure 3.11: Performance of the predictive model based (A) on substitutions at all 76 identified locations, and (B) on the 61 locations in the RBS region only. Each graph shows predictive performance of the model using the set of HI assay titres used in a previous study (Liao et al., 2008), and also when using an extended set, incorporating assay results from additional studies in order to obtain greater precision. When used with the model developed for this study, the improved precision of the extended set leads to improved performance.

Both models provide relatively consistent performance across a broad range of cell sizes in which the residue density varied considerably. This indicates robustness in the underlying approach. The falloff in the predictive power in the validation set at cell sizes below 12Å in the 76-location model and 8Å in the RBS-only 61-residue model suggests that, at these small cell sizes, the model is over-fitting to the training set by overweighting those locations that are significant in that set. The RBS-only model provides significantly improved results in the cell size range of 6Å to 10Å without significant degradation at a larger cell size. While this could be caused by over-fitting in the 76-location model, the result is interesting in view of the discovery by Okada et al. (2010) of wild-type antibodies binding in the mid region which are not haemagglutinin inhibiting, as such antibodies would be overlooked by the HI assay.

Table 3.5 compares the performance of the models discussed in this section with the best results obtained by Liao et al. (2008) and with the best results from our previous work, demonstrating that the identification of immunoactive sites through discovery of antigenic clusters can be used to overcome the weak specificity of our previous approach.

Model	Sensitivity (%)	Specificity (%)	MCC
Complete 76 locations – average scores	87.5	79.9	0.57
Complete 76 locations – best result	90.0	89.5	0.73
RBS only – average scores	86.5	80.6	0.60
RBS only – best result	90.0	92.1	0.77
Multiple regression, GM4 (Liao et al., 2008)	83.2	93.5	n/a
Banded model M4 (Lees, Moss, and Shepherd, 2010)	97	57	0.55

Table 3.5: Comparison of the sensitivity and specificity of models based on identified cluster participants when tested against the extended HI titre data set compared to the best models from two previous studies. For the models developed for this study, both the average figures (averaged across 8 cell sizes from 8Å to 22Å) and the best result obtained (at 12Å for the complete 76 location model and 8Å for the RBS only model) are shown.

3.9 Sensitivity of the Model to Grid Orientation

While the employment in the analysis above of grids with multiple cell sizes provides some assurance that the results are not overly biased by selection of a favourable grid configuration, I also examined sensitivity to grid orientation by changing the grid origin. In this analysis, the origin was progressively shifted along the x-, y- and z- axes simultaneously (i.e., along the major diagonal) by an increasing percentage of the cell size. The MCC value (averaged across 8 cell sizes from 8Å to 22Å) was determined at each orientation (Figure 3.12). The results confirm a lack of bias to grid configuration (overall mean MCC 0.57, standard deviation 0.04).

Figure 3.12: Mean Matthews correlation coefficient obtained by the predictive model based on substitutions at 76 locations and cubical cell sizes ranging from 8Å to 22Å in 2Å steps. The x, y, and z co-ordinates of the origin of the grid were progressively shifted by a percentage of the cell size, with '0% grid offset' coinciding with the grid origin used in section 3.10.

3.10 Alternative Predictive Models Considered

Liao et al. (2008) considered several predictive methods, reporting the best performance with a multiple-regression approach in which substitutions at roughly 20 locations were selected by the model during the analysis of the training set. To attempt to reduce noise induced by substitutions unlikely to prove effective, amino acids were categorised into groups using one of six different approaches of increasing complexity (the simplest having three groups: polar, non-polar and charged; and the most complex distinguishing additionally between aliphatic and aromatic non-polar amino acids and introducing additional groups for cysteine and glycine), with only substitutions that caused a switch of groups being considered significant. With the models developed for this study, I found that grouping approaches such as those reported by Liao et al., or the introduction of additional explanatory variables based on properties such as charge or hydrophobicity, did not improve performance significantly. I believe that this reflects the relatively low level of noise in difference scores at the locations selected.

3.11 Discussion

A key challenge in building predictive models of antigenic escape is to separate those amino acid substitutions related to escape from those that occur for other reasons. The 'antigenic clustering' approach addresses this challenge by seeking patches of multiple substitutions that are spatially co-ordinated and occur together as part of a known antigenic transition. The technique is similar to that used by earlier researchers to infer the binding location of monoclonal antibodies by studying substitutions in escape mutants (Wiley, Wilson, and Skehel,

1981), but here it is applied computationally to wild-type strains, introducing the requirement for spatial co-ordination to eliminate at least some of the substitutions arising from other causes.

The extent to which the antigenic clusters identified match genuine epitopes could only be established categorically via experimentation. At the time of writing there has only been limited study of the epitopes of human antibodies, and studies have focussed on unusual and rare properties, such as broad-spectrum binding, rather than reviewing what is typical in the population as a whole. In this chapter I have compared in some detail the results of one experimental study which is particularly relevant to this work, and in which I have established several correspondences. Beyond experimental evidence, some other factors give confidence. Firstly, the antigenic clusters identified lie within the range of identified H3N2 antigenic sites. Secondly, the set of 76 locations identified as participating in clusters is to a large extent a superset of sets of locations other researchers have identified as having undergone periods of selective pressure, or determining membership of antigenic clusters. Finally, I have demonstrated that counts of substitutions at the 76 locations can be used in predictive models, and can overcome the problems of specificity associated with our previous models, which were based on larger numbers of locations.

Given that the survey of antigenic clusters utilised strains isolated across a period of 34 years, models based on substitutions at the identified locations are likely to have greater generality than those based on much smaller numbers of locations (Liao et al., 2008; Huang and Yang, 2011), which have been necessary until now in order to obtain acceptable predictive performance.

While examples of atypical antibody binding exist (Kubota-Koketsu et al., 2009; Barbey-Martin et al., 2002), my results lend support to the dogma that anti-HA1 antibodies generally bind in the region of the five antigenic sites. At first sight, the restriction to these sites is surprising, particularly as the antibodies used in the experiments from which the sites were deduced were of mouse origin rather than human. Indeed, it is perhaps surprising that one observes a general consistency in antigenic escape across a population, as evidenced by the consistency of HA assay results discussed in the previous chapter. The underlying reasons are likely to be a combination of genetic and structural factors. While V-region gene rearrangement and ensuing secondary diversification allow for a very large gene repertoire, there is evidence that the typical antibody response to influenza can, in practice, be more limited in diversity (Kavaler et al., 1990; Clarke et al., 1990; Lingwood et al., 2012). That conformational epitopes themselves have structural and compositional features, making some parts of a protein more immunogenic than others, is discussed in Rubinstein et al. (2008).

This analysis suggests that, in H3N2 in particular, antibody binding to HA1 takes place in two distinct regions: the RBS region and the mid region, with evolution and escape taking place more rapidly in the former than the latter region. This led us to speculate that the clustered nature of the H3N2 antigenic map observed by Smith et al. (2004) might be caused by interplay between antigenic escape in the two regions. This possibility is explored in Chapter 4. It should be noted that an epitope is sufficiently large to span multiple antigenic sites. The RBS region includes sites A and B, while the mid region includes sites C and E. Substitutions in site D are present in both RBS and mid region clusters.

The analysis in this chapter, and that of the experimental study quoted, are both confined to the HA1 chain. One antibody identified in the experimental study, which was found to bind in the mid region, was neutralising but not haemagglutination inhibiting. This suggests an orientation remote from the RBS, and a neutralisation action through fusion inhibition rather than steric interference with receptor binding, both factors which would make it likely that the epitope would extend into the HA2 chain, given its location and its role in membrane fusion. The ready isolation of such an antibody from a human volunteer would have important implications for vaccine selection and epidemic forecasting: the human stalk-binding antibodies that have been the subject of X-ray crystallographic studies were specifically selected for their broad-spectrum affinity and are, by contrast, extremely rare: Corti et al., for example, isolated ‘a few’ broadly binding plasma cell derivatives from a single donor, having screened 104,000 cells from 8 donors. This possibility – and the possibility that other mid-range clusters might also have epitopes extending into HA2 – led us to examine variation in HA2 in more detail. This work is described in Chapter 5

.

4 Exploring the clustered behaviour of antigenic distance in Influenza A H3N2 Haemagglutinin

4.1 Introduction and Motivation

The antigenic evolution of influenza A H3N2 Haemagglutinin, as revealed by antibody binding assays, is known to be punctuated. When antigenic distances between H3N2 strains are used to construct two dimensional antigenic maps (see the following section for a discussion of antigenic mapping), this behaviour gives rise to a clustered behaviour (Figure 4.1).

(this figure is not included in the public version)

Figure 4.1: Antigenic Map of selected H3N2 strains showing clustered behaviour (Smith et al., 2004). The position of each strain is shown in colour, antisera are shown in outline. The size and shape of each data point represents uncertainty in the calculated position. Strains are assigned to clusters in the map by k-means clustering. Clusters are named after a vaccine strain considered representative of the cluster in terms of its position and isolation year. Each square represents one unit of antigenic distance.

Studies of conformational antibody binding in general have indicated that, of the 15-22 residues typically comprising a B-cell epitope, a small subset of 5-6 contribute most of the binding energy (Laver et al., 1990). In antibody/haemagglutinin complexes, cases have been found in which mutation of just 1 or 2 residues will give rise to antigenic escape (Smith et al., 2004; Jin et al., 2005).

In Chapter 3, we saw that antigenic evolution in HA1 in wild type human H3N2 strains is associated with mutations that cluster in two regions: one region surrounding the RBS at the membrane-distal end of the haemagglutinin stalk, and the second (the 'mid region') lying closer to the HA1/HA2 interface. These regions coincide with, but are more restricted than, the antigenic sites A-E identified in *in vitro* studies based on monoclonal mouse antibodies (Wilson and Cox, 1990).

I was interested to explore the potential of these two phenomena (variation in binding strength between residues in an epitope, and multiple regions associated with antigenic evolution) to give rise to punctuated evolution of antigenic distance. To facilitate this, I developed a model to simulate antigenic evolution.

4.1.1 Antigenic Maps

Edelstein and Rosen (1978) introduced the concept of *shape space* as a means to model protein-protein interaction. The nature of the two interacting surfaces, for example those of an epitope and paratope, are considered to be represented by two co-ordinates in multidimensional space, where the dimensions describe both the geometric shape of the surfaces, and also the other physical and chemical characteristics relevant to binding. The molecular affinity observed experimentally is assumed to be a monotonic function of the distance between the two co-ordinates in shape space.

Perelson and Oster (1979) used experimental data to estimate the dimensionality of shape space, finding it to be approximately five-dimensional. Lapedes and Farber (2001) applied ordinal multi-dimensional scaling (MDS) to antigenic distances derived from several small sets of influenza HI assay data. They confirmed that this data could be represented accurately in five dimensions, but no fewer. An important result from this work was the discovery that logarithmic antigenic distance is linearly related to distance in shape space.

Smith et al. (2004) used metric MDS to model a large set of H3N2 HI data, covering the period 1968 – 2002. Interestingly, in view of the above results, they found that antigenic distances could be recalled from the model without appreciable loss of precision even if the shape space was restricted to two dimensions, thus allowing for the construction of conveniently viewable antigenic maps, such as the map illustrated above in Figure 4.1.

A complication arises from the definition of antigenic distance, which is normally defined as the logarithm of the ratio of a homologous titre to a heterologous titre (Section 2.2.1). In some, relatively rare, cases, the heterologous titre exceeds the homologous titre, leading to a negative

antigenic distance under the above definition. In other cases (as was the case with some experimental data used by Smith et al.), homologous titre values are not available. Smith et al. adopted a modified definition of antigenic distance, in which the numerator is the highest observed titre in that column of the assay table, rather than the homologous value. Another complication seen in HI data is the reporting of threshold values, often denoted by '<' in the assay table. Here the antibody reactivity was too low to be observed within the limits of the assay. The reported result is a lower bound on antigenic distance rather than a specific value.

MDS algorithms are typically based on matrix transformations: however this approach requires a complete matrix (Borg and Groenen, 1997). Matrices of antigenic data tend to be sparse, particularly when, as in the case of Smith et al., many assay tables are brought together in order to create a dataset covering an extensive period.

Smith et al. derived co-ordinates in the antigenic map by minimising an error function using conjugant gradient optimization (Flannery, Teukolsky, and Vetterling, 1988). In the case of numerical assay values, the error function took the form

$$e(D_{ij}, d_{ij}) = (D_{ij} - d_{ij})^2$$

where D_{ij} is the antigenic distance between antigen i and antiserum j as measured in the assay, and d_{ij} is the Euclidean distance as represented in the map. In the case of threshold values, the function is modified to

$$e(D_{ij}, d_{ij}) = (D_{ij} - d_{ij})^2 g(D_{ij} - 1 - d_{ij})$$

where

$$g(x) = \frac{1}{(1 + e^{-10x})}$$

and D_{ij} is the reported lower bound on antigenic distance. In this modified function, $g(x)$ is a 'squashing function' which ensures that the reported error is close to zero when $d_{ij} > D_{ij}$.

Smith et al. tested the effectiveness of their maps by utilising them to predict HI values that were missing from the data used to construct the maps, and then determining those values experimentally. They noted that prediction error decreased 'only slightly' as the number of dimensions was increased from 2 to 5. This result may reflect noise or lack of precision in the data set rather than inherent low dimensionality of shape space.

Prediction of missing data could be beneficial to surveillance, allowing the number of assays to be reduced, or the number of reported strains to be increased. Cai et al. (2010) noted that this problem of data completion is similar that set in the Netflix Challenge: a problem that received much attention as a result of the large prize offered by Netflix Corporation for a satisfactory solution (Bennett and Lanning, 2007). They developed an approach to predicting HI assay data termed MC-MDS, which brings together these advances in matrix completion with MDS.

Cai et al. suggested also that novel emerging strains would tend to be antigenically different to all of those observed previously in the population, over some window roughly coincident with herd immunogenic memory. They introduced a further term into the MDS error function to account for this tendency. This phenomenon of 'temporal bias' essentially introduces a preferred direction of evolution into the map. Such a preferred direction is seen in assay data when comparing an antiserum with a strain whose isolation date differs by up to 5-10 years: antigenic distance tends to increase over time. Beyond that point, antibody reactivity is typically too low for an antigenic distance to be determined, and whether temporal bias exists beyond that point is therefore not deducible from the data. In my simulation I examine the effect of preferred directionality in antigenic maps in one simulation run (see Figure 4.9), but otherwise do not include a directional preference. I demonstrate that introducing a biologically reasonable degree of preference does not impact the overall conclusions.

4.1.2 Determination of Cluster Quality

In developing the antigenic map shown in Figure 4.1, Smith et al. (2004) used *k*-means clustering (MacQueen, 1967) to identify clusters. *k* represents the number of clusters into which the data set should be divided, and must be provided as an input to the algorithm. The algorithm also requires an initial assignment of the data to clusters (which may be random). The algorithm proceeds through successive optimisation steps, in each step determining the centroid of each cluster (the *means*), and then assigning each data point to its closest mean. The algorithm terminates when no assignments change during an optimisation step.

While this method is widely used, the requirement for the user to provide a value for the parameter *k* has disadvantages. An optimal value of *k* has to be identified, typically by running the algorithm with various values of *k*, so that the best value can be selected. The results of *k*-means clustering also depend critically on the initial assignment to clusters: ideally, therefore, the algorithm should also be run with multiple initial assignments. Antigenic map simulation experiments require the creation of a large number of antigenic maps. The need to run the clustering algorithm multiple times on each one would therefore introduce significant overhead.

The selection of a 'best' value of *k* and the best set of input conditions can be made by reviewing the resulting clusters, either by eye, or algorithmically. In initial simulations using k-means clustering, I found that the act of partitioning a map into a certain number of coloured clusters had a powerful influence on perception, making it hard to determine objectively the optimum number to use: a computational assessment is therefore to be preferred. The value of *k* in the map shown in Figure 4.1 was set to 11. The determination of that value is not discussed by the authors.

Computational approaches for determination of the best *k* include evaluation of the average Silhouette (Rousseeuw, 1987), and information theoretic methods (Sugar and James, 2003). The selection of suitable criteria raises an important point for this work: while most observers would agree that the map shown in Figure 4.1 displays a degree of clustering, what exactly do we mean by clustering in this context, and can the 'degree' to which a map is clustered be quantified (Jain, 2010)?

In the case of antigenic maps, the key quality exhibited by a clustered map is that of density: the surprising observation is that antigenic distance evolves discontinuously rather than continuously, so that there are regions of low density in the map. Algorithms have been developed to derive clusters directly from density analysis: in particular DBSCAN (Ester et al., 1996) and its derivatives.

The DBSCAN algorithm requires two initial parameters: Ɛ, and MinPts. Ɛ is a distance, defining the extent of a neighbourhood around a point. In a two-dimensional map such as ours, the neighbourhood is typically defined as a circle with radius Ɛ. MinPts represents the minimum number of points required to form a cluster. The algorithm starts by considering the neighbourhood around an arbitrarily selected data point. If at least MinPts data points are found in the neighbourhood, a cluster is started, otherwise any points in the neighbourhood are provisionally marked as 'noise'. If a cluster is started, each point assigned to the cluster is then examined in turn, and the cluster widened to incorporate that point's neighbourhood, provided that the MinPts criterion is met for the neighbourhood. The cluster widening continues as long as suitable data points are identified. Once all suitable points have been processed, an as-yet unvisited point is selected for examination. The algorithm terminates once all points have been visited.

Density-based algorithms have limitations in handling high-dimensional data (Jain, 2010), and DBSCAN will not work well with clusters of widely varying density (an issue that addressed in a later algorithm, GDBSCAN (Jörg et al., 1998)). Neither of these limitations raise concerns in

this context. More complex variants are available which will infer a cluster hierarchy, but such a hierarchy is not relevant to this work.

For this work, DBSCAN was selected for cluster determination. The advantages of this algorithm over *k*-means clustering are:

- The density-based search algorithm fits naturally with the concept of antigenic clustering;
- In contrast to *k*, the parameters Ɛ and MinPts do not require tuning for individual maps in a series;
- It will classify outlying points as noise rather than allowing them to influence cluster partitioning;
- It is simple to implement;
- It is determinate: there is no dependence upon an initial cluster assignment.

Use of DBSCAN requires values to be selected for Ɛ and MinPts. Empirically, I selected Ɛ=0.10 and MinPts=3 as suitable parameter values for the simulated antigenic maps generated in this study (Figure 4.2). Sensitivity to these settings is discussed later.

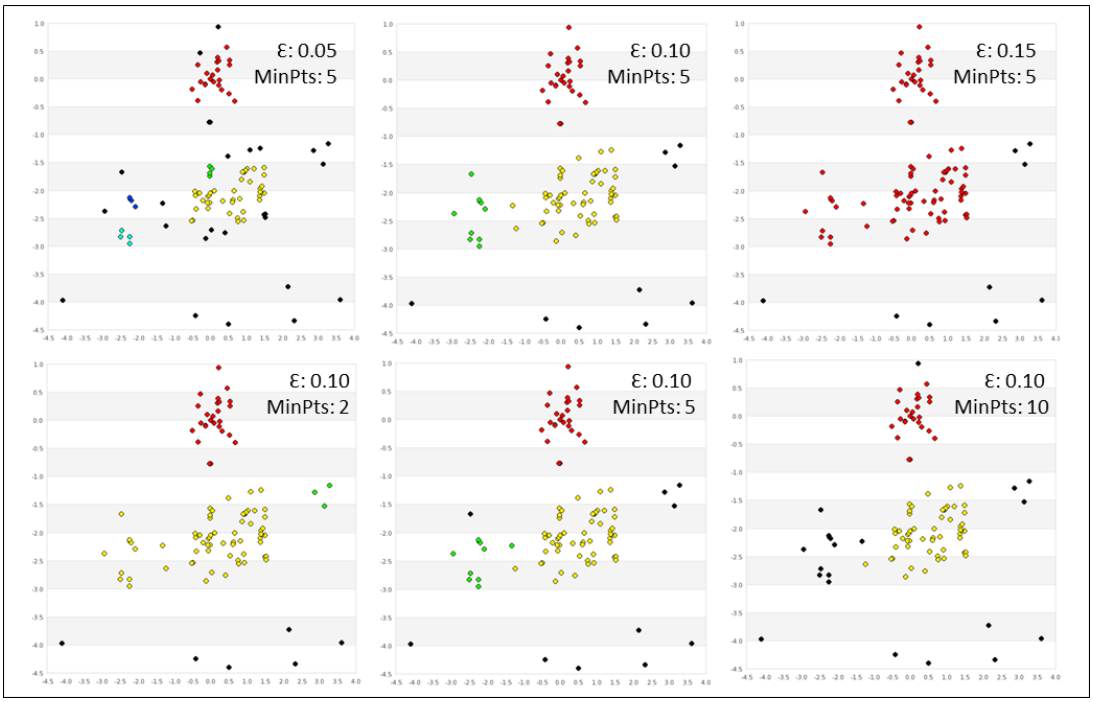


Figure 4.2: Clusters derived by DBSCAN for selected values of Ɛ and MinPts. Ɛ effectively defines the minimum boundary between clusters. Selecting a low value of Ɛ leads to the identification of many small clusters, while a high value will cause cluster boundaries to become overlooked. A value in the region of 0.10 is reasonable for this data (top row). The selection of MinPts is less sensitive: low values may allow noise to blur the boundary between clusters, while high values will cause small clusters to be classed as noise. A value of approximately 5 appears reasonable for this data (bottom row). The scales on x and y axes show deviation (in antigenic units) from the initial strain.

4.1.3 Measurement of the 'Degree of Clustering'

In order to compare maps produced with different parameter settings, we require a metric of cluster quality. For this purpose I employed Silhouette scores (Rousseeuw, 1987). Silhouette scores have performed well in evaluation by a number of researchers against other measures such as the Dunn index (Dunn, 1973), the Davies-Bouldin index (Davies and Bouldin, 1979), and the Rand index (Rand, 1971) as a reliable measure of cluster quality (Clifford et al., 2011; Grafahrend-Belau et al., 2008; Lovmar et al., 2005; Pearson et al., 2007) and have been used in a wide variety of contexts.

The Silhouette score for a single point *i* is calculated from two metrics *a(i)* and *b(i)*. *a(i)* is a measure of the average dissimilarity of point *i* to other points in the same cluster. While any metric of dissimilarity may be used, in this case I use the average Euclidean distance between point *i* and the other points in the cluster. *b(i)* is the lowest average dissimilarity of point *i* to points in another cluster: the average dissimilarity to each cluster (apart from that cluster including point *i*) is calculated and the lowest value is taken. Again, although any measure may be used, I use the average Euclidean distance as the measure of dissimilarity. From *a(i)* and *b(i)*, the Silhouette score for point *i*, *s(i)*, is calculated as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

and the Silhouette score *S* for the entire map is the mean of *s(i)* for all points in the map that are cluster members (in other words, excluding points classified by the clustering algorithm as noise).

From the definition, it will be seen that values of *s(i)* can range between -1 and 1. Values close to 1 indicate strong clustering, while values close to 0 indicate little clustering. Negative values indicate an erroneous assignment of points to clusters, and are not found in simulated antigenic maps when classified by DBSCAN. In the degenerate case of just a single cluster, I assign the map a Silhouette score of 0. Figure 4.3 shows Silhouette scores obtained for a number of simulated antigenic maps.

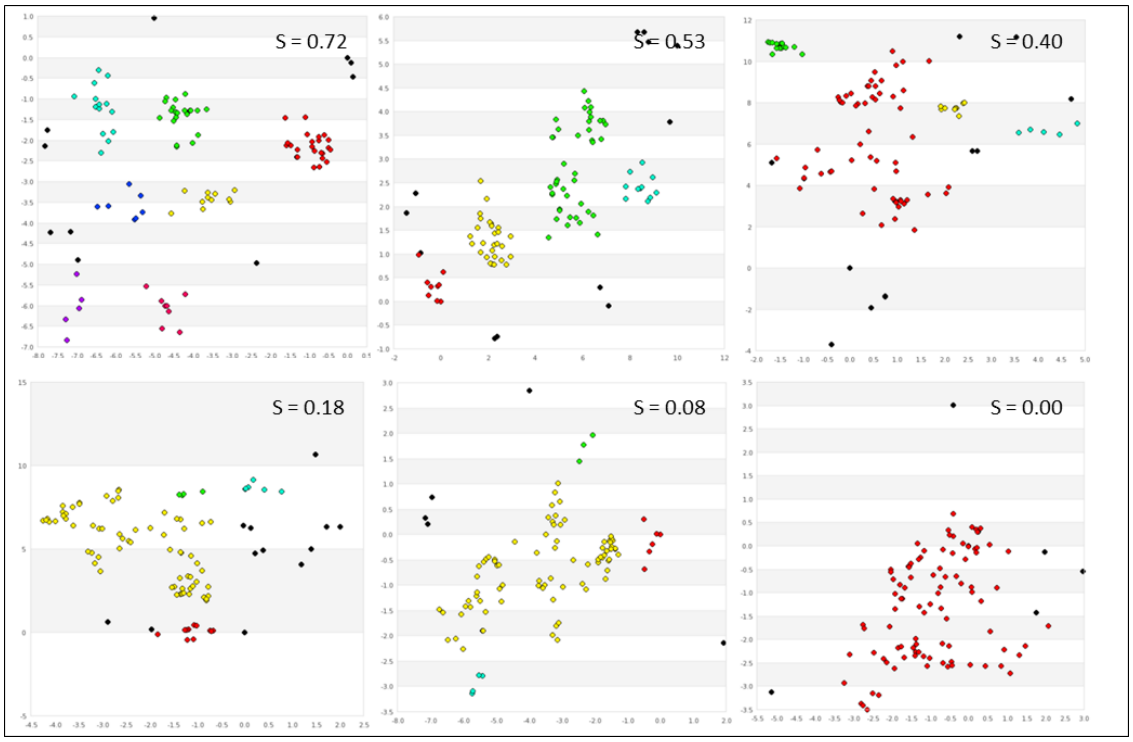


Figure 4.3: Simulated antigenic maps created with a range of model parameters in order to demonstrate the Silhouette score S as a measure of cluster quality. In these maps, and other maps in this chapter, each cluster is shown in a separate colour, and data points classified as noise are shown in black.

4.1.4 Model Principles

The model presented in this chapter is intended to be a simple simulation that incorporates just sufficient functionality to simulate the two phenomena of interest. The model portrays antigenic evolution on a two dimensional antigenic map. It starts with a single strain, which is evolved through a number of generations. In each generation, there is a chance of antigenic evolution. If antigenic evolution occurs, a new strain is created at a random distance from its parent within a predefined range, at a random bearing on the map (the random bearing reflects the lack of temporal bias in the model, discussed in the previous section). The newly spawned strains are considered fitter than their parents with respect to the host population, as a result of the antigenic distance, and, over a number of generations, ancestor strains are retired from the population once newer strains exist that are separated by more than a threshold antigenic distance.

To simulate the significant change in antibody binding energy occasionally provoked by a mutation at one of small number of locations, the model provides a chance for a mutation to cause an elevated antigenic distance, over and above the normal predefined range. This elevated distance is associated with the relatively rare substitution of one of these key residues.

To simulate the contribution from mutations in two independent regions, the model provides for two chances of antigenic evolution in each active strain per generation, with the antigenic parameters for the two chances being set separately. In this way, I can model the relatively lower mutability of the mid region compared to the RBS region and can determine the effect of greater or lower contributions to antigenicity from the two regions.

4.1.5 Model Parameters

The following parameters, for each site, can be set by the experimenter:

p_{AD} - The probability that an antigenically different strain will be created in a generation;

$\Delta_{min}, \Delta_{max}$ – The new strain will be created with an antigenic distance of at least Δ_{min} from the parent strain, and, unless the antigenic distance is elevated by the mutation of a key residue, a distance no greater than Δ_{max};

p_{high} - The probability that an antigenically different strain will have an elevated antigenic distance, caused by the mutation of a key residue;

h - The multiplier to be applied in the case of an elevated antigenic distance;

In each generation, each extant strain is provided with the opportunity to create antigenically dissimilar new strains. The antigenic distance Δ_{AD} between the parent strain and the new strain is determined as follows. For each of the two sites, four random numbers, r_g, r_l, r_h and r_d are generated. Each of these can take a value from 0 to 1.

If $r_g \geq p_{AD}$, no new strain is created.

If $r_g < p_{AD}$ and $r_h \geq p_{high}$, a new strain is created, and

$$\Delta_{AD} = \Delta_{min} + r_l(\Delta_{max} - \Delta_{min})$$

If $r_g < p_{AD}$ and $r_h < p_{high}$, a new strain is created, and

$$\Delta_{AD} = \Delta_{min} + r_l(\Delta_{max} - \Delta_{min}) + h\Delta_{max} + r_d - 0.5$$

To simulate the two independent sites, in each generation, I allow for the creation of two strains at independent antigenic distances Δ_{AD1} and Δ_{AD2}, using independently generated random numbers and independent parameters. The two sets of parameters are distinguished by using a numbered suffix, e.g. r_{g1}, p_{AD2}. Representative values are discussed in Section 4.2.1.

In addition to the above, the model requires values for three values which apply to strains created at either site:

Fitness Threshold – The antigenic distance that must separate two strains in order for the newer to be considered fitter.

Parent Survival Time – The number of generations that a prior strain should survive a newer strain that is at a sufficient antigenic distance to be considered fitter.

Target Strains – The number of child strains to be generated in this simulation.

Pseudocode for the simulation is given below:

```
Create a single strain at co-ordinates (0,0), mark as 'active'
LOOP until we have created the required number of strains
  FOR EACH active strain
    FOR EACH site in the strain
      assign random values between 0 and 1 to rG, rL, rH, rD
      IF rG < pAD THEN
        IF rH >= pHigh THEN
          deltaAD = deltaMin + rL*(deltaMax-deltaMin)
        ELSE
          deltaAD = deltaMin + rL*(deltaMax-deltaMin)
                      +h*deltaMax + rD - 0.5
        END IF
        create new active strain at a distance deltaAD and at a random bearing
      END IF
    END FOR
  END FOR

  FOR EACH active strain
    IF a more recent strain exists
      AND distance between them > fitnessThreshold THEN
      decrement strain lifetime
      IF strain lifetime <= 0 THEN
        mark strain as inactive
      END IF
    END IF
  END FOR
END LOOP
```

4.1.6 Parameter Values

In this report I present a range of results to illustrate model sensitivity to particular parameters, but I regard the following values as being representative of current experimental data. In the simulations to follow, one parameter is generally varied while others are held at the values listed in this section.

$p_{AD1} = 0.5,\ p_{AD2} = 0.25$

The model is relatively insensitive to the absolute value of the mutation chance, as this only affects the parent survival time relative to the development of new strains. In the previous chapter, I found that transitions in the RBS region were roughly correlated with the need to update the H3N2 vaccine strain, which occurred 16 times between 1972 and 2005. I found transitions in the mid region roughly correlated with antigenic cluster transitions, of which there were 9 in the same period, providing a ratio between the two of approximately 1:2.

$\Delta_{min} = 0,\ \Delta_{max} = 0.75, h = 3.3$

Smith et al (2004) found that a single residue mutation caused on average 0.37 units of antigenic change. A single mutation in a Beijing/1992-like strain was found to cause 2.5 units of change. The Δ_{min} and Δ_{max} settings provide a range around the average value, and h extends this to the level found in the exceptional single mutation.

$p_{high} = 0.125$

16 HA1 residues differ between A/Panama/2007/1999 and A/Wyoming/03/2003. Of these, the mutations at two locations were found to cause significant antigenic drift with mutations at other locations being relatively insignificant (Jin et al., 2005). The value of 0.125 (1/8) reflects this distribution.

Fitness Threshold=0.2, Parent Survival Time=1

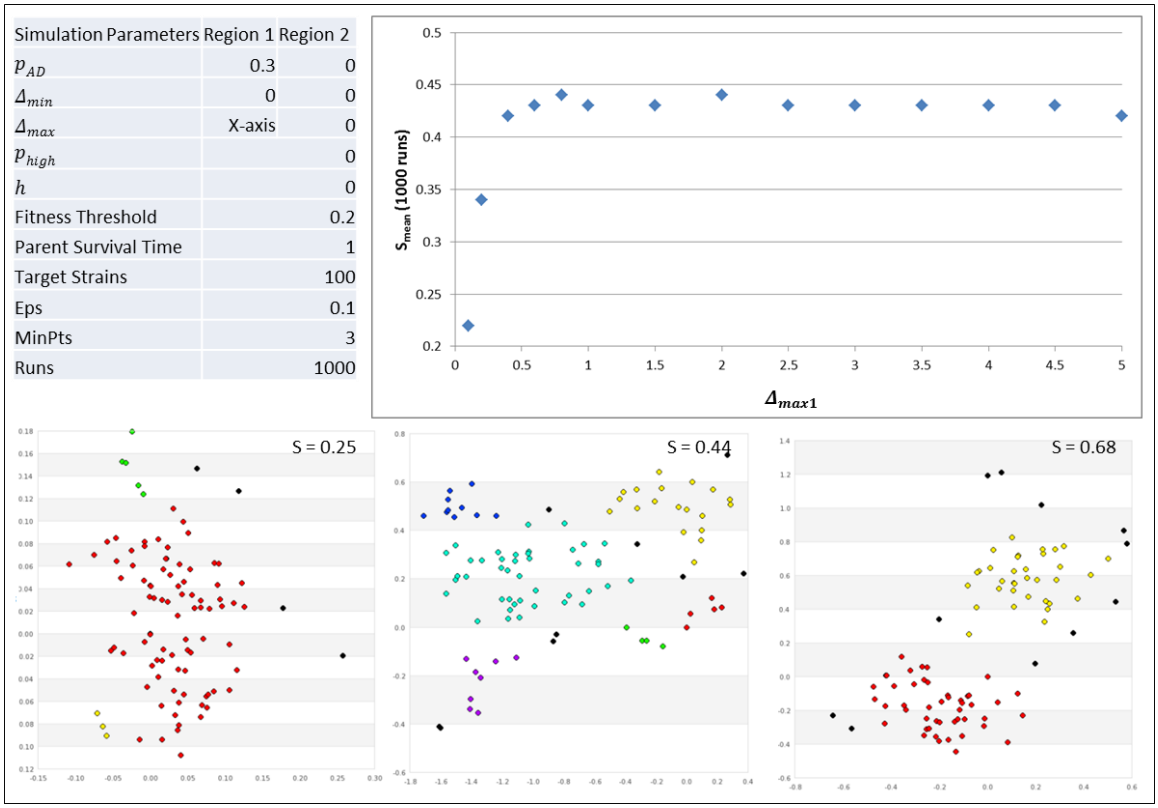
Varying Fitness Threshold and Parent Survival Time did not alter the character of the models unless they were set sufficiently high to swamp the results with numerous extant strains. The values are set to reflect the fact that, in general, there is just one dominant strain of H3N2 at any time.

4.2 Implementation

The antigenic map simulator, including DBSCAN for cluster determination and Silhouette for cluster quality evaluation, was implemented in PHP on the web site described in Appendix B. Simulations can be run interactively from the web site. Further PHP code (not accessible from the web site) was developed to carry out long-running simulations with varying simulation parameter values.

4.3 Model Results

With a single active mutation region and $p_{high} = 0$, cluster quality is generally low and does not vary with Δ_{max}, except at very low values, comparable to the value of the Fitness Distance (Figure 4.4). At such low values of Δ_{max}, 'parent' strains are less likely to be removed from



circulation, altering the pattern of antigenic development.

Figure 4.4: Results with a single mutable region and $p_{high} = 0$ show antigenic movement but little differentiation into clusters. The graph shows variation of S (averaged over 1000 simulations, SE < 0.01) with Δ_{max}. The plots show typical simulated maps obtained with the settings shown. These illustrations of typical maps and those in subsequent figures were created by selecting parameter values from the graph likely to give the required value of *S*.

If $p_{high} > 0$, a higher cluster quality is achieved, with the quality increasing with increased h (Figure 4.5). Clusters obtained typically show much greater differentiation than those obtained with $p_{high} = 0$, even at a comparable S_{mean}.

Simulation Parameters	Region 1	Region 2
p_{AD}	0.3	0
Δ_{min}	0	0
Δ_{max}	1	0
p_{high}		0
h		X-axis
Fitness Threshold		0.2
Parent Survival Time		1
Target Strains		100
Eps		0.1
MinPts		3
Runs		1000

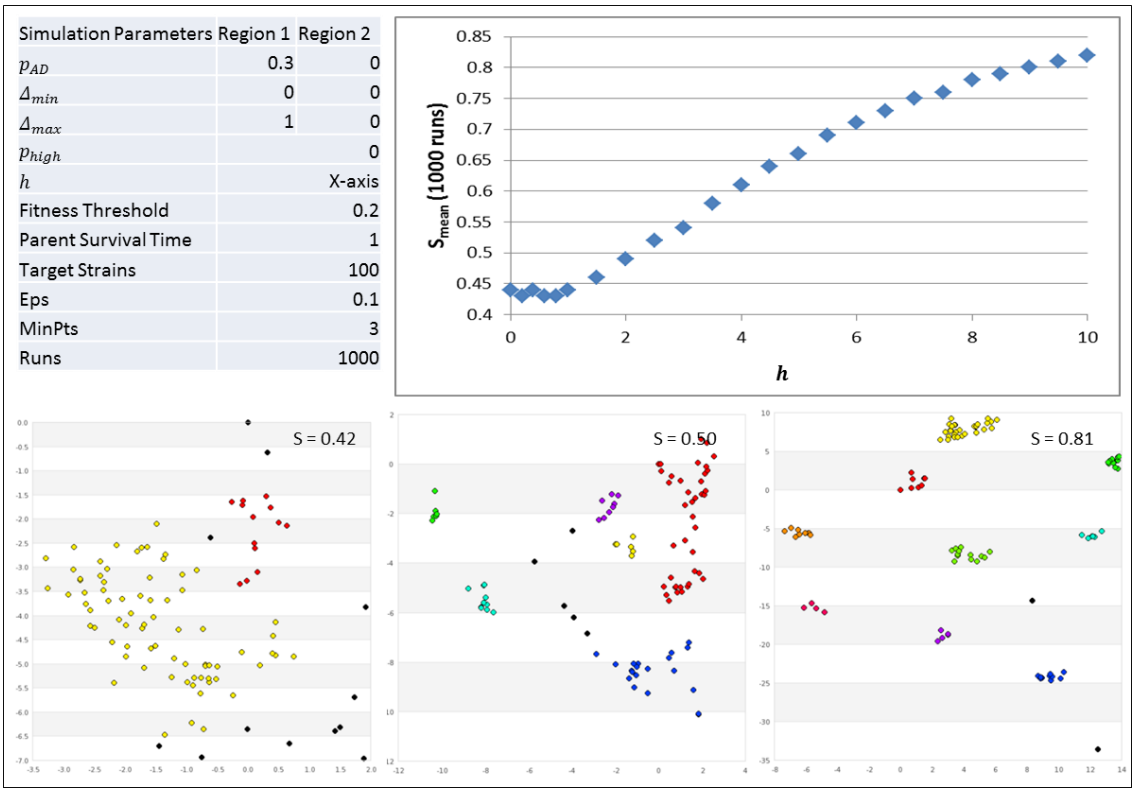


Figure 4.5: With a single mutable region, the degree of clustering increases as h is increased. For values at or below 1, there is effectively no 'exceptionally elevated' contribution and cluster quality remains low. Results for Smean are averaged over 1000 simulations, SE < 0.01.

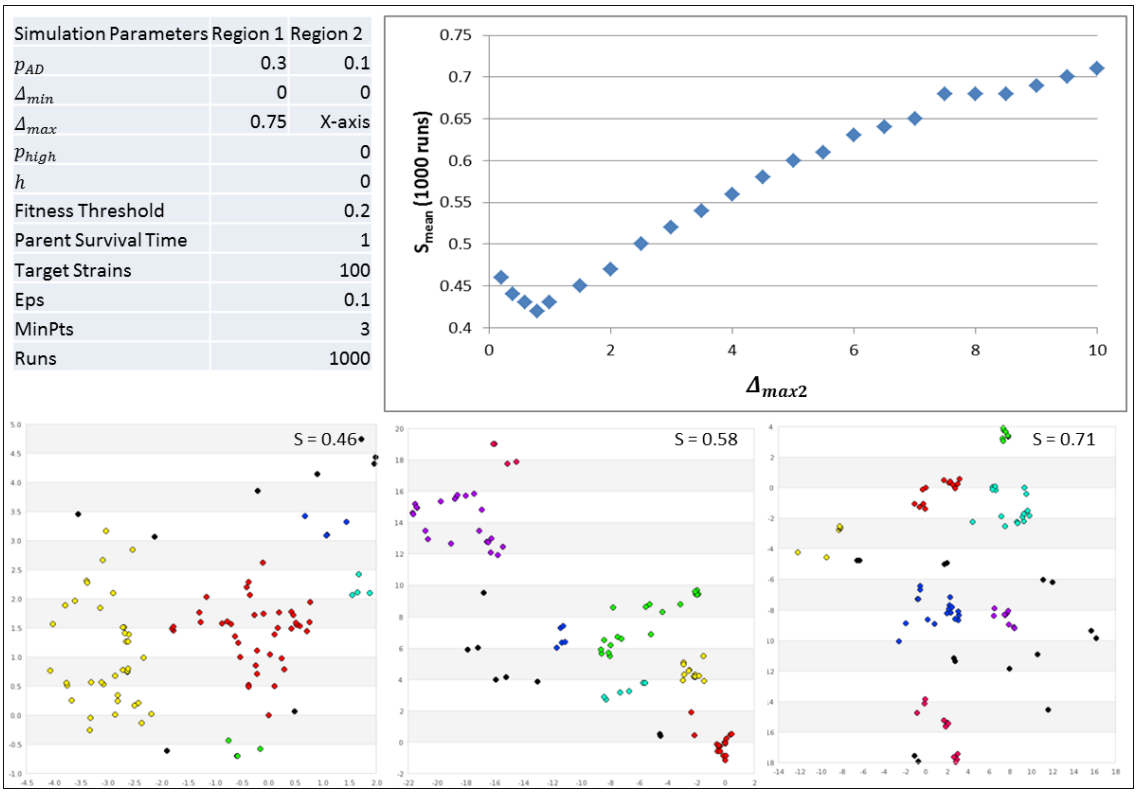


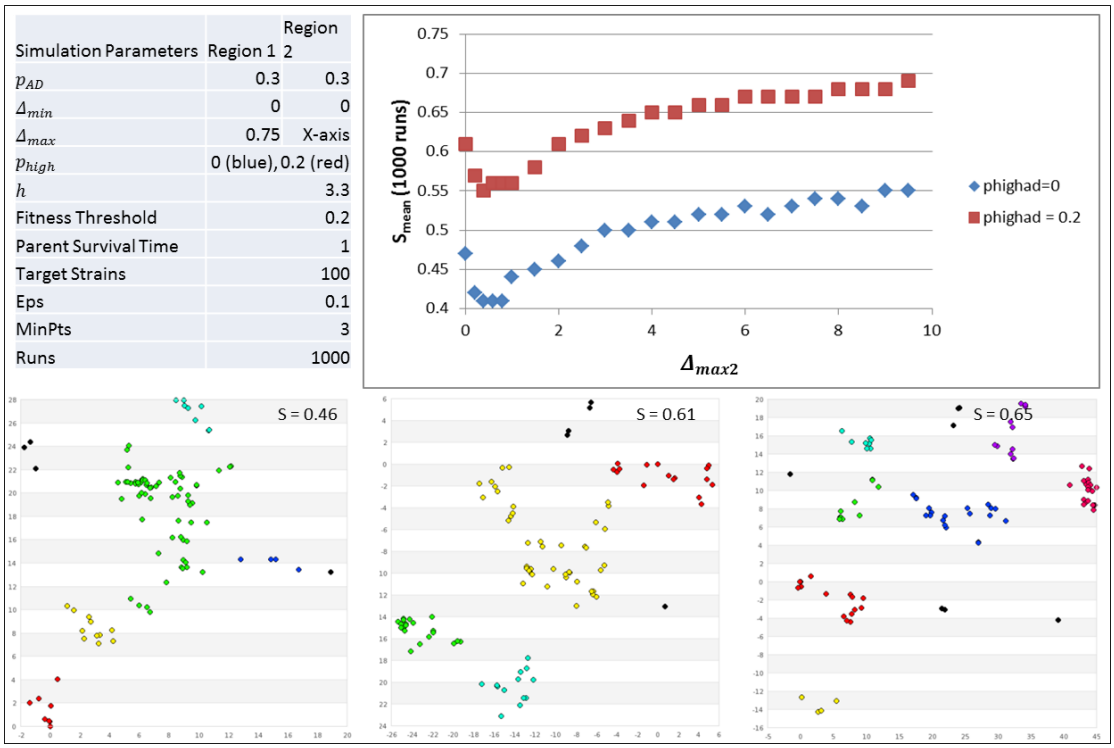
Figure 4.6: With two mutable regions and $p_{high1} = p_{high2} = 0$, cluster quality increases as Δ_{max1} and Δ_{max2} diverge. A higher cluster quality is observed if the mutation chance p_{AD} differs in the two regions (compare the maximum value of Smean achieved in this figure (0.7) with that achieved in Figure 4.4 (0.55): in the latter case p_{AD} was set to 0.3 in both regions. Results for Smean are averaged over 1000 simulations, SE < 0.01.

When $p_{high} = 0$, cluster differentiation is observed with two active mutation regions provided Δ_{AD1} differs from Δ_{AD2}. Cluster quality increases as the values diverge (Figure 4.6). Cluster quality is higher if the mutation chance p_{AD} of the two regions differ (this can be seen by comparing the graph in Figure 4.6 with the lower line in Figure 4.7: the probabilities differ in the former case and are equal in the second, while other parameters are identical).

From Figures 4.5 - 4.7, we can see that both the introduction of a chance of a nonzero p_{high} and the introduction of a second active region can elicit high quality clustering that is not observed in the absence of both factors. The effect is cumulative: with two active regions, introducing a nonzero p_{high} increases cluster quality (Figure 4.7).

Figure 4.8 shows the distribution of *S* values obtained from runs of 1000 simulations at typical settings with two active regions and $p_{high} > 0$. If, with reference to Figure 4.2, we define the threshold of acceptable cluster quality to be $S = 0.5$ in order to consider maps of a similar cluster quality, then at $S_{mean} = 0.6$, 80% of simulations yield a score at or above the threshold, while at $S_{mean} = 0.7$, the proportion rises to 92%. It should be noted that cluster quality does not of itself

have biological significance. Here we are interested in the proportion of maps that have at least



a similar quality to that seen in Figure 3.5.

Figure 4.7: With two mutable regions, introducing a nonzero p_{high} increases cluster quality.

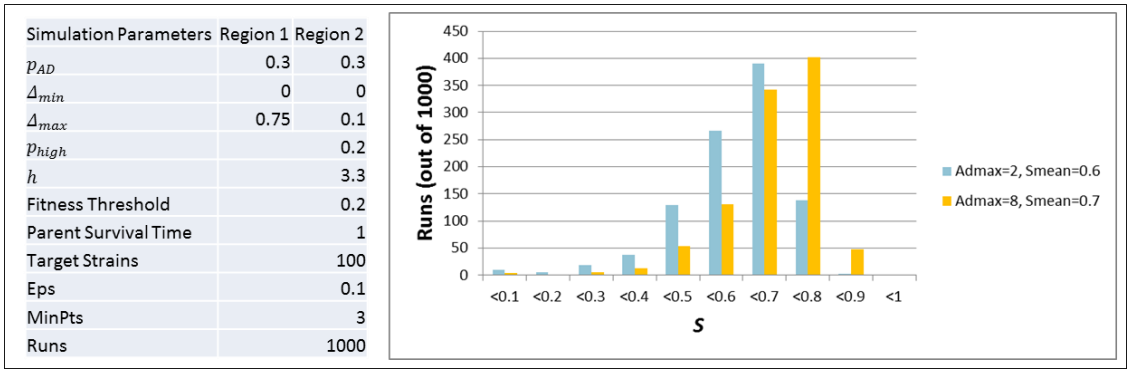
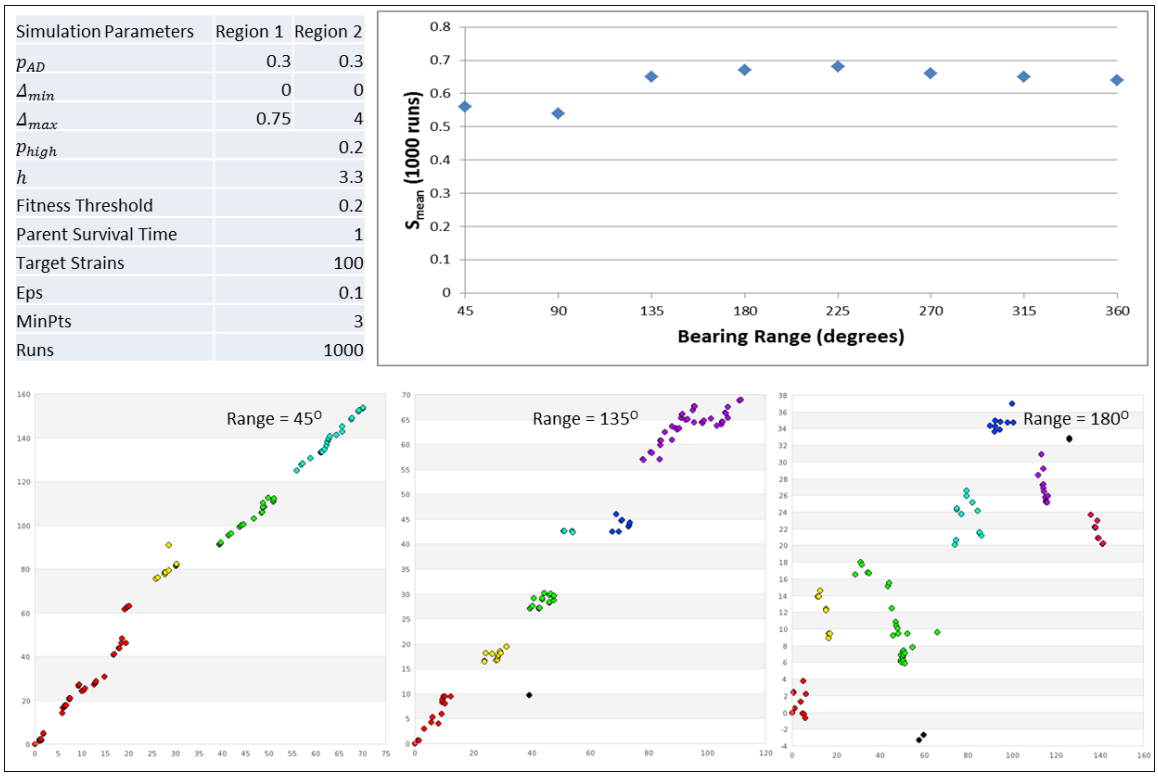


Figure 4.8: Distribution of S values obtained from runs of 1000 simulations employing two active regions, using parameter values selected from those of Figure 4.7 to provide Smean values of 0.6 and 0.7. Results for Smean are averaged over 1000 simulations, SE < 0.01.

In the simulations used above, a new and antigenically different strain is introduced at a random bearing from its 'parent' strain. As discussed in Section 4.1.1, an element of directional bias might more closely reflect antigenic evolution, in that, at least over a short to medium generational window, a strain is likely to be antigenically different to all its successors. I

examined the impact of directional bias on cluster quality by restricting the range of the bearing of child strains (Figure 4.9). With two active regions, and parameter settings that achieve S_{mean} = 0.65, bearing ranges of 135° or higher produce a comparable S_{mean} and comparable antigenic maps. Smaller bearing ranges produce antigenic maps with distinctly pronounced linearity, and somewhat lower values of S_{mean}, probably as a result of the elongated, somewhat one-



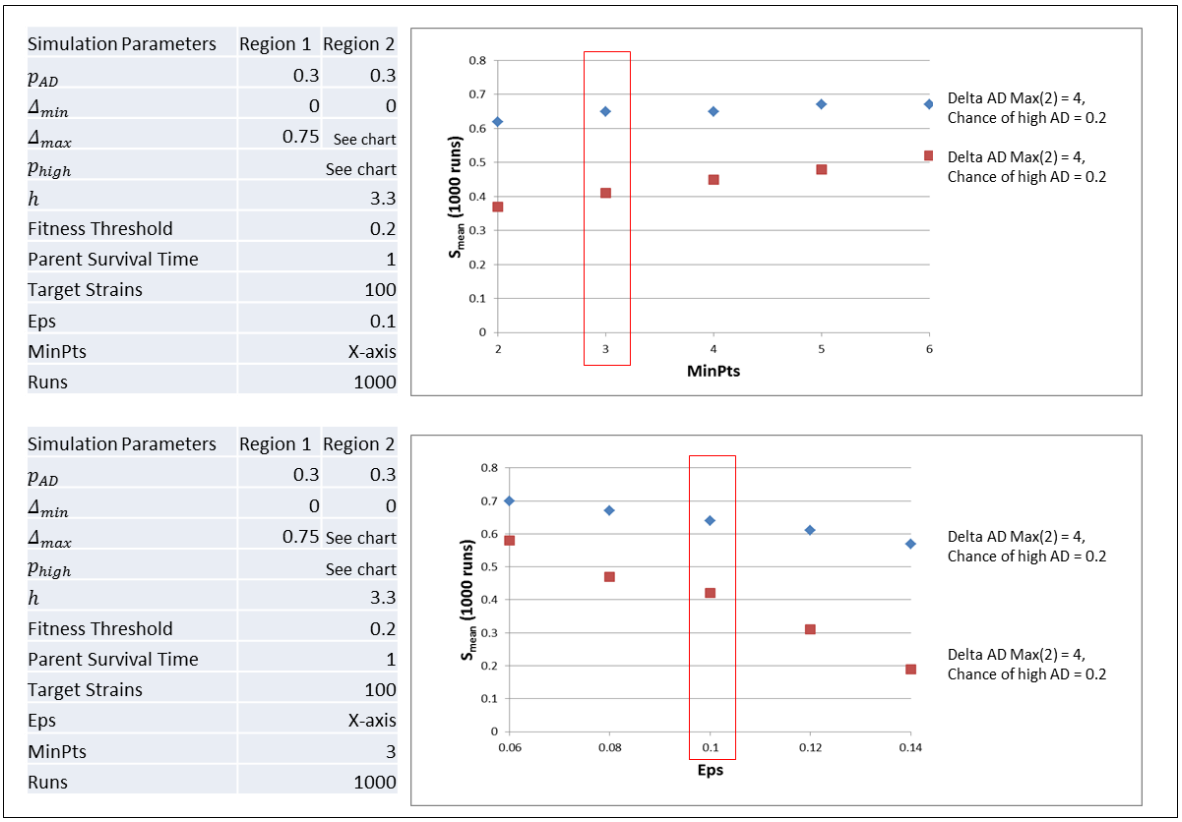
directional, clusters.

Figure 4.9: Restricting the bearing of new strains from their parents has little effect on the Smean obtained at typical parameter settings taken from those of Figure 4.6, until the bearing range is reduced below 135o. Below that value, antigenic maps take on a linear character, and the achieved Smean is reduced. Results for Smean are averaged over 1000 simulations, SE < 0.01.

Figure 4.10 shows the sensitivity of the achieved S_{mean} value to variations in the DBSCAN parameters MinPts and Ɛ. Overall, these demonstrate a linear trend, in which a clear distinction is preserved between the achieved S_{mean} for a high quality and a low quality cluster: hence the analysis above is not unduly sensitive to the selected values. As MinPts increases, the number of points classed as 'noise' increases, and this increases the quality of the residual clusters. This has a greater impact on low scoring clusters: hence at high MinPts, the scores start to converge. As Ɛ increases, DBSCAN will require points to be further apart in order to be sorted into separate clusters. For a given map, this will result in larger clusters and hence reduce overall

cluster quality. The impact is greater for lower quality clusters, as these will tend to be separated by a smaller distance.

Figure 4.2 demonstrated that the values we selected for these parameters provide good results with few outliers classed as 'noise'. From Figure 4.10, we can see that electing other reasonable values for these two parameters would alter the precise values obtained in the analysis but



would not alter the overall qualitative conclusions.

Figure 4.10: Sensitivity of Smean to variation in the DBSCAN parameters MinPts and Eps (Ɛ) from the values used in this analysis (bracketed in red). Results for Smean are averaged over 1000 simulations, SE < 0.01.

4.4 Discussion

In this analysis, I have demonstrated the applicability to antigenic maps of DBSCAN as a clustering algorithm, and of Silhouette as a cluster quality score. These tools could prove useful in other antigenic mapping applications, and either separately or in combination could be used to provide a firmer foundation for the division into clusters. On the basis of density distribution, some cluster boundaries in the H3N2 map shown in Figure 4.1, such as that between BK79 and

SI87, are questionable: density mapping may identify more biologically relevant clusters than *k*-means clustering.

The simulations suggest that the influence of some mutations with exceptionally high antigenic effects can produce the clustered behaviour seen in H3N2 antigenic maps. The existence of such mutations has been demonstrated experimentally, but it is not known whether mutations of sufficient strength occur frequently enough to explain the observed antigenic cluster transitions as there are examples where the binding energy is relatively evenly distributed (Lee and Air, 2002; Venkatramani et al., 2006). I have demonstrated that another factor, the interplay between antigenic activity at two independent regions, could also produce antigenic clusters, and have shown in simulation that it is possible for the two phenomena to work in combination.

If clustering activity is indeed caused by one or both of the above phenomena, we might expect to see periods in the H3N2 antigenic evolution when such activity is at a reduced level, for example when antigenic activity is focussed on one site, and the antibodies involved have relatively evenly distributed binding energy across their epitopes. More generally, we might expect clustering activity to fluctuate as a function of these effects. Figure 4.1 does suggest a fluctuation in clustering activity – for example the clusters for WU95 and SY97 are clearly separated, while those for SI87 and BE89 are much less so – but such observations could easily be affected by sample bias.

5 Antigenic Escape in Influenza A HA2

5.1 Introduction and Motivation

In Chapter 3, I developed techniques for predicting HA antibody binding locations in wild-type human H3N2 strains. The analysis was confined to HA1, which is typically considered to include the antigenically active regions of the protein.

The identification of several predicted antibody binding locations in the 'mid' region, a region somewhat distant from the tip, and a region in which HA1 overlaps with HA2, suggested the possibility of HA2 participation in B-cell epitopes. Also, although first reported some years previously (Okuno et al., 1993), this work coincided with the announcement of a number of broad-spectrum antibodies binding in the HA2 stalk (Table 5.1). As well as understanding whether HA2 residues could contribute to epitopes in the mid region, I therefore also wished to apply the techniques to an analysis of the stalk.

Reference	**Subtypes Bound**	**Epitope determined by**
Okuno et al., (1993)	H1, H2	Generation of escape mutants
Throsby et al. (2008)	H1, H2, H5, H6, H8, H9	Inference, substitutions, modelling
Sui et al., (2009)	H1, H2, H5, H6, H9, H11, H13, H16	Crystallography
Ekiert et al., (2009)	H1, H5	Crystallography
Wang et al., (2010)	H1, H3, H5	Ab induced via vaccination with HA2-based protein
Wei et al., (2010)	H1, H2, H5	Competition assay with stem mutant
Kashyap et al., (2010)	H1, H5	Prediction based on results of Sui et al. and Ekiert et al.
Clementi et al., (2011)	H1, H2, H5, H9, some H3	Alanine mutagenesis
Ekiert et al., (2011)	H3, H7, H10, H15	Crystallography
Corti et al., (2011)	H1, H2, H3, H4, H5, H6, H7, H9, H10, H13	Crystallography
Dreyfus et al., (2012)	H1, H2, H3, H4, H5, H6, H7, H9, H10, H12, H13, H14, H15, H16, type B	Crystallography

Table 5.1: Studies of stalk-binding antibodies discussed in this chapter.

A broad-spectrum vaccine is a key objective of influenza vaccine research, and the induction of broadly-binding anti-stalk antibodies has been advocated as a possible approach (Wang and Palese, 2009). Such an approach would depend on the stalk being unable (or slow) to achieve antigenic escape. To date, stalk-binding antibodies isolated for structural study have been specifically selected for their reactivity against a broad range of strains. The stalk has a low rate of evolution compared to other regions of HA, in particular the region close to the RBS. Nevertheless, as we shall see later in this chapter, fixations and periods of polymorphism do

occur in the stalk, and escape mutants to neutralising stalk binding antibodies have been cultured *in vitro*. Studies of recently isolated stalk-binding antibodies are necessarily retrospective. The extent to which broad-binding properties may be preserved across future strains, and, in particular, whether evolutionary pressure might elicit escape following antibody emergence, is the subject of this chapter.

5.2 Availability of HA2 Sequences

The sequence database contains full length sequences for approximately 1,300 H1N1 strains and 3,000 H3N2 strains isolated between 1968 and 2008 (Figure 5.1).

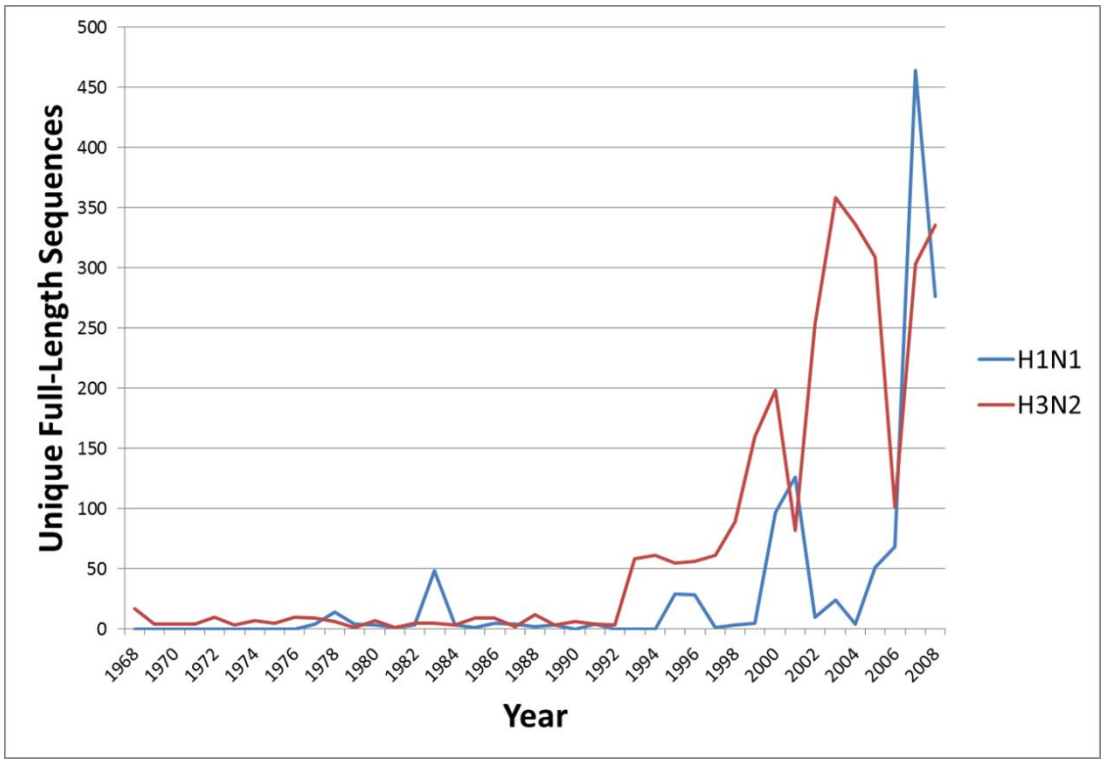


Figure 5.1: Number of unique strains isolated between 1968 and 2008 for which full-length sequences exist in the database.

The H3N2 record provides a more consistent set of samples over a longer period – not only because H1N1 was not in circulation until 1977, but also because a higher number of sequences are available in most years between 1977 and 2000: 823 unique full-length sequences for H3N2 strains are available in that period, compared to 262 H1N1 sequences.

To examine the contribution of HA2 residues to clusters identified in Chapter 3, we require full length HA sequences (i.e., including HA2 in its entirety) for the dominant strains used in that analysis as listed in Figures 3.2-3.4. As some strains were missing from the database, I reviewed publicly available sources including GenBank (http://www.ncbi.nlm.nih.gov/GenBank), Uniprot

(http://www.uniprot.org/), NCBI's Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html), the Influenza Research Database (http://www.fludb.org) and the GISAID database (http://www.gisaid.org) in order to obtain as many as possible. Although sequencing efforts are reducing the number of gaps, some required full length HA sequences are not available. I therefore constructed 'hybrid' sequences for these strains by combining the consensus HA1 sequence of the strain with the consensus HA2 sequence taken from the full sequence strain differing from the target strain in the lowest number of HA1 residue positions (Table 5.2).

Required Strain	**Selected HA2 hybrid strain**	**Differing HA1 locations**
A/Texas/1/1977 (H3N2)	A/Memphis/1/1977 (H3N2)	3
A/Mississippi/1/1985 (H3N2)	A/Hong Kong/14/1983 (H3N2)	3
A/Sichuan/2/1987 (H3N2)	A/Hong Kong/7/1987 (H3N2)	2
A/Beijing/352/1989 (H3N2)	A/Beijing/353/1989 (H3N2)	1
A/Wuhan/359/1995 (H3N2)	A/New York/631/1996 (H3N2)	2
A/Hong Kong/1550/2002 (H3N2)	A/Hong Kong/CUHK23162/2002 (H3N2)	2
A/Fujian/411/2002 (H3N2)	A/Hong Kong/CUHK5316/2003 (H3N2)	0
A/Bayern/7/1995 (H1N1)	A/New York/643/1995 (H1N1)	1

Table 5.2: Dominant strains required for cluster analysis for which no full-length sequence could be obtained, and the full-length strain selected from the database for hybridisation. The third column shows the number of locations in HA1 at which the hybrid strain's residue differs from that of the required dominant strain.

To provide an indication of likely accuracy of these hybrid sequences, I analysed all full-length sequences whose HA1 sequence differed from one of the strains in this survey with a known HA2 in three or less locations (three being the highest difference amongst the hybrid candidates). The analysis suggests that, on average, the 'hybrid' HA2 may be expected to differ from the desired HA2 in one position or less, and that a difference in two positions is unlikely (Table 5.3).

Target Strain	Number of full length strains with HA1 varying in 3 locations or less	Number of such strains with an HA2 sequence differing from the target	Total number of residues at variance	Maximum residues at variance in any one strain	Average residues at variance per strain
A/Hong Kong/1/1968	24	8	8	1	0.33
A/Port Chalmers/1/1973	5	0	0	0	0.00
A/England/42/1972	5	0	0	0	0.00
A/Victoria/3/1975	0	-	-	-	-
A/Bangkok/1/1979	5	0	0	0	0.00
A/Philippines/2/1982	1	1	1	1	1.00
A/Leningrad/360/1986	3	3	3	1	1.00
A/Beijing/353/1989	9	3	3	1	0.33
A/Beijing/32/1992	0	-	-	-	-
A/Johannesburg/33/1994	16	0	0	0	0.00
A/Sydney/5/1997	1	0	0	0	0.00
A/Moscow/10/1999	25	8	12	2	0.48
A/Wellington/1/2004	291	228	254	3	0.87
A/California/7/2004	131	14	14	1	0.11
A/Wisconsin/67/2005	2	0	0	0	0.00
A/Perth/16/2009	205	28	29	2	0.14
(TOTAL)	*723*	*293*	*324*	*3*	*0.30*

Table 5.3: Analysis of sequences with an HA1 which is close to that of required H3N2 strains, showing the variation of that HA2 of those strains compared to the known HA2 of the required strain. Across all 723 strains in this sample, only 1 differed from the target HA2 sequence in more than two locations.

5.3 HA2 Participation in Mid Region Clusters

H1N1

The clusters inferred in two of the transitions shown in Figure 3.2 are changed as a result of the inclusion of HA2 substitutions. In the transition A/Brazil/11/1978-A/Chile/1/1983, a substitution at HA2 57 is added to the mid region cluster (Figure 5.2A). In the transition A/Chile/1/1983-/Singapore/6/1986, HA2 residues 57 and 77 are added, as a result of which the mid cluster moves lower and the fixation at HA1 58 is dropped (Figure 5.2B).

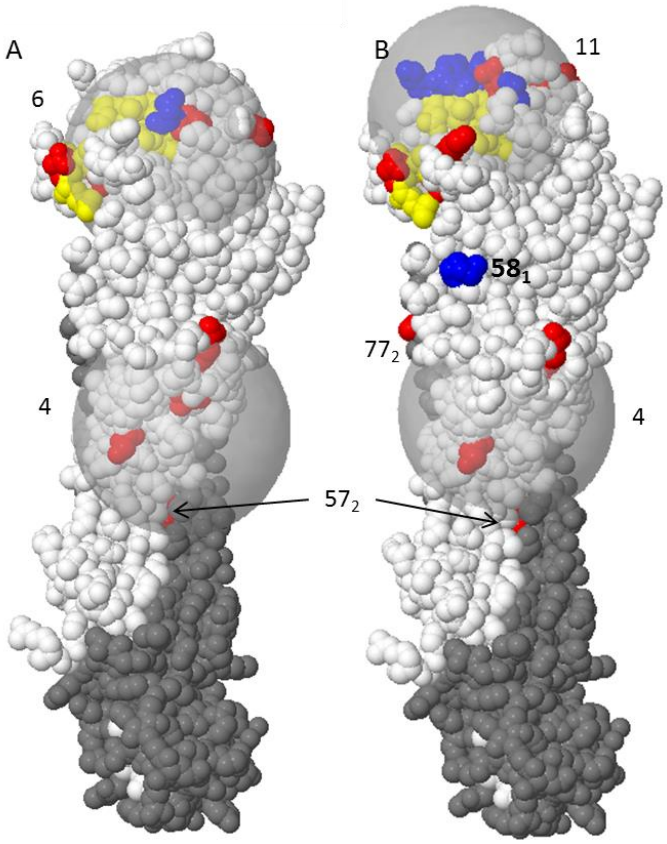


Figure 5.2: Clusters in the H1N1 series which are altered as a result of considering HA2 substitutions. A - A/Brazil/11/1978-A/Chile/1/1983. B - A/Chile/1/1983-/Singapore/6/1986. As in Figures 3.2-3.4, each cartoon indicates positions on a single HA monomer at which substitutions occur between two dominant strains. Those substitutions that become fixed in the population are shown in blue; others in red. The RBS is coloured yellow. Other HA1 residues are white, and other HA2 residues grey. The number of substitutions in each cluster is shown as a single number. The location code of substitutions of interest is given as a number followed by the subscript 1 or 2 to indicate HA1 or HA2 respectively.

H3N2

Clusters deduced for three transitions in the H3N2 series outlined in Figures 3.3 and 3.4 are affected by consideration of HA2 substitutions. In the transition A/Beijing/32/1992-A/Johannesburg/33/1994 (Figure 3.3), a new mid region cluster is formed, incorporating HA2 56 and two HA1 substitutions (Figure 5.3A).

In A/Johannesburg/33/1994-A/Wuhan/359/1995 (Figure 3.3), HA2 46 and 56 are included in the mid region cluster, which moves closer to the membrane as a result, and a new cluster of three residues is inferred on the HA head, away from the RBS (Figure 5.3B). HA1 124 is within 35Å of either cluster on the head, but is assigned to the cluster closest to the RBS as the algorithm seeks the largest possible clusters.

In A/Bangkok/1/1979-A/Sichuan/2/1987 (Figure 3.4), a new mid cluster is formed from HA2 55 and 66, and HA1 307. All three residues are buried in the trimeric structure, making it unlikely that this inferred cluster is indicative of antigenic escape (Figure 5.3C).

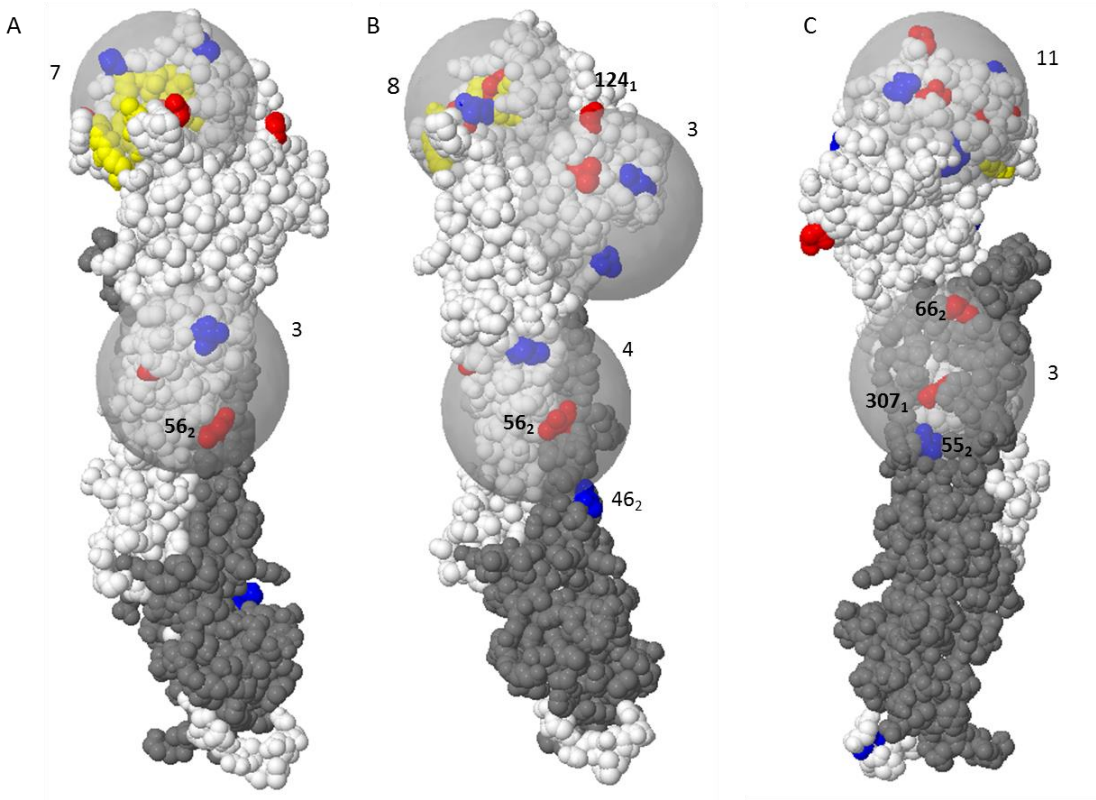


Figure 5.3: Clusters in the H3N2 series which are altered as a result of considering HA2 substitutions. A - A/Beijing/32/1992-A/Johannesburg/33/1994. B - A/Johannesburg/33/1994->A/Wuhan/359/1995. C - A/Bangkok/1/1979-A/Sichuan/2/1987 (showing the side of the monomer facing in to the trimer). Colouring and legend as in Figure 5.2.

Conclusion

Excluding the additional mid region cluster inferred in the transition A/Bangkok/1/1979-A/Sichuan/2/1987, in which all the residues are buried, and which therefore does not seem indicative of antigenic escape, HA2 substitutions play a part in just two H1N1 and two H3N2 transitions. Only one includes a fixation in an HA2 residue. Participation of HA2 residues in mid region clusters may occur, but it is at most limited, and consideration of HA2 residues does not significantly change the number or size of inferred clusters in the analysis of Chapter 3.

5.4 The Action and Activity of Stalk Binding Antibodies

From analysis of the binding activity of panel of seven monoclonal antibodies to proteolytic fragments of HA2, Varecková et al. (2003) identified four linear peptide regions on HA2 to which antibodies bound. Antibodies binding to three of these regions provide protection against

lethal challenge in mice, but only antibodies binding close to the N-terminus of HA2, which is known to be associated with the pH-induced structural transformation that occurs during the process of membrane fusion (Skehel and Wiley, 2000), are virus neutralising (Varecková et al., 2003; Staneková et al., 2012).

Among the neutralising anti-HA2 antibodies, very broad-spectrum examples are rare (Corti et al., 2011), but studies suggest that less broad-spectrum types are widely present in the adult human population (Sui et al., 2011; Staneková et al., 2012). Hu et al. (2013) reported the isolation of stalk-binding neutralising antibodies from a 2009 pandemic H1N1 vaccine recipient. Wrammert et al. (2011) found that, in 9 individuals infected with the H1N1 pandemic strain, a majority of neutralising antibodies were stalk-binding and broadly cross-reactive. They contrast this with the vaccination response to annual strains, in which the proportion of such antibodies is much lower. Staneková et al. (2012) report a negligible increase in anti-HA2 antibody levels after conventional vaccination, compared to an approximately twofold increase following a natural infection with a seasonal H3N2 strain: hence a comparison of antibody levels in natural infection on one hand and in vaccination on the other may be misleading. Although researchers have hypothesised that stalk-binding antibodies are induced by the appearance of novel strains (Palese and Wang, 2011; Chiu et al., 2013), results to date do not provide a conclusive picture of their prevalence in seasonal infections.

Results from several studies suggest that antibodies binding in the stalk may be more frequent in those subjects that have been subjected to repeated infections. Children have a restricted antibody response to HA compared to adults, with the response typically focusing in the area of the RBS (Natali, Oxford, and Schild, 1981; Nakajima, Nobusawa, and Nakajima, 2000). Repeated exposure of mice to different influenza strains can induce a broader response including the elicitation of stalk-binding antibodies (Vanlandschoot et al., 1998; Wang et al., 2010; Krammer et al., 2012). The global influenza surveillance program, which focuses in large part on the immune response of serologically naïve animals, may therefore underestimate the role of such antibodies in conferring protection to the adult human population. Until recently, the programme has also relied almost exclusively upon the HI assay, which is not sensitive to the neutralisation mechanism of stalk-binding antibodies.

Stalk-binding antibodies specific to group 1 influenza subtypes isolated to date, as well as the group 1/group 2/Influenza B antibody CR9114, have a common germline descent from the variable heavy chain gene IGHV1-69 (Throsby et al., 2008; Sui et al., 2009; Dreyfus et al., 2012), possibly as a result of the unusual ability of antibodies descended from this germline to bind to conserved hydrophobic pockets (Sui et al., 2009).

Concerns have been raised that the stalk region of HA on the intact virion may not be readily accessible to antibodies: however a recent study using cryoelectron microscopy to examine H1N1pdm virions incubated with C179 IgG antibody found that 75% of the HA trimers were complexed with at least one antibody particle, with a random distribution of bound and unbound trimers on the viral surface (Harris et al., 2013). A modelling study based on measurement of single HA fusion events suggests that six HA trimers must undergo the fusogenic transition in order to achieve membrane fusion (Blumenthal et al., 1996). On that basis, the antibody density found by Harris et al. would be sufficient to inhibit membrane fusion.

5.5 The Limits of Broad-Spectrum Stalk Binding Antibodies

The ability or otherwise of the broad spectrum antibodies cited in Table 5.1, which are known to be rare, to neutralise any strain in the group(s) to which they are specific is of interest. If they are indeed completely effective against a subtype, we would not expect them to elicit mutations in that subtype, since no escape from their action would be possible. If, on the other hand, escape or partial escape by certain strains is possible despite the structural constraints exerted by the fusogenic function, then there is the possibility of antigenic evolution, and the possibility that an antigenic region will develop in the locality of their epitopes. In this section I examine the arguments and analysis presented by researchers of such antibodies (Table 5.4), in particular of those antibodies that feature in crystal studies, as those studies provide the most specific information relating to epitopes and therefore allow for in-depth analysis.

In the studies considered in this section, the case for broad-spectrum binding is generally developed along the following lines:

- Assays are used to demonstrate neutralising action or binding strength of the antibody to various strains (typically one strain of each subtype in the relevant groups);
- The epitope (i.e., the precise atomic contacts) is determined from crystal studies of the affected subtypes;
- The conservation of each epitopic residue is analysed;
- Follow-up assays are conducted of wild-type amino acid variants revealed by the conservation analysis that were not covered by the first set of assays;
- Escape mutants are cultured, both to identify critical residues and also to assess the likelihood of escape in the wild.

We will address each topic in turn, looking both at the evidence presented in published studies and also at its limitations.

Antibody	PDB Structure(s)	Reference
C179	None	Okuno et al., (1993)
CR6261	3GBN (H1) 3GBM (H5)	Throsby et al. (2008), Ekiert et al., (2009)
D8 F10 A66	3FKU (F10, H5)	Sui et al., (2009)
7A7 12D1 39A4	None	Wang et al., (2010)
PN-SIA28	None	Clementi et al., (2011)
CR8020	3SDY (H3)	Ekiert et al., (2011)
FI6	3ZTN (H1) 3ZTJ (H3)	Corti et al., (2011)
CR9114	4FQI (H5) 4FQY (H3) 4FQV (H7)	Dreyfus et al., (2012)

Table 5.4: Broad-spectrum antibodies whose descriptions are considered in this section.

Assays Employed In Binding Studies

As the neutralising action of the antibodies in question is related to fusogenic activity rather than the inhibition of cell attachment, the HI assay cannot be used to quantify their efficiency. Neutralisation assays, accompanied by assays of antibody binding strength, are typically employed. Neutralisation assays determine the titre at which an antibody will reduce viral plaque formation to a given percentage of that achieved by a control. When conducted with human or animal serum, they are generally regarded as representative of viral activity *in vivo*, as evidenced by their presence in recent WHO surveillance reports. The concentrations at which stem-binding antibodies might be naturally present in human serum are not well understood, although Sui et al. (2011) estimate that they are present in concentrations of up to 0.1μg/ml. The best (lowest) IC_{50} neutralising concentrations obtained from assays of stem binding antibodies are between 0.1 and 1.0 μg/ml: however many subtypes to which broad-spectrum antibodies bind only neutralise at much higher concentrations (Table 5.5). This suggests that the neutralising action of broad-spectrum stalk-binding antibodies against some cited strains and subtypes is weak at physiological levels. In addition, one study utilised artificial cell constructs, which may not accurately mimic natural conditions, such as the dense packing of haemagglutinin on the viral membrane and could therefore exaggerate neutralising activity (Wang et al., 2011).

Neutralisation assays were conducted with whole antibody IgG with the exception of antibody F10, for which Fab fragments were used. Lee at al. (2012) found that the bivalent binding of IgG to an RBS epitope, in which the immunoglobulin is thought to span two HA trimers, can

increase apparent affinity compared to a monovalent Fab fragment by roughly three orders of magnitude and reduce neutralization IC_{50} concentrations by roughly two orders of magnitude. The same effect may apply to stalk binding IgG. On the other hand, the more demanding packing requirements close to the viral membrane, which were noted in the previous section as resulting in only partial occupancy of stalk epitopes, may reduce or eliminate the effect: Ekiert et al. (2009) note that Fab fragments of antibody CR6261 neutralise as potently as intact IgG.

Antibody	IC_{50} < 10 μg/ml	IC_{50} > 10 μg/ml
CR6261	H1, H5, H6, H8, H9	H2
CR8020	H3	H7
7A7	H3	
12D1	H3	
39A4	H3	
A06*	H1pdm	H1, H5
PN-SIA28	H1, H2, H5, H9	H3, H3
FI6		H1, H3
CR9114	H1, H2, H4, H8, H9, H12	H3, H5, H6, H7, H10, H11, H14
$F10^{+}$	H5	

Table 5.5: IC50 neutralisation concentrations for each subtype neutralised by a stalk-bonding antibody, for those antibodies and subtypes for which figures are available. Where more than one strain of a subtype was tested, the highest IC50 concentration is used. The physiological concentration is estimated to be approximately 0.1 μg/ml, hence for those subtypes in the second column in particular, the naturally occurring antibody may not be fully neutralising in vivo. Figures are taken from the references cited in Table 5.4. Figures for CR6261 are from Throsby et al. *Figures for A06 are minimal inhibitory concentrations. $^{+}$Antibodies tested are whole antibody IgG apart from F10 which is represented as a Fab fragment.

Determination of the Epitope

The specific residues involved in interactions between HA and antibody are inferred from crystal structures by using software to predict hydrogen-bonds and van der Waals interactions. The results obtained are subject to error in the crystal determination, variation in the crystal structure compared to the shape under physiological conditions, and also depend upon the software and parameters employed for the calculation (which are rarely stated). My own analyses (presented in Chapter 3) are in general agreement with experimental reports, but in most cases I identify some participating residues that are not identified in the study, and vice versa. In all the studies reviewed, researchers identified, from a single analysis, a single set of epitopic residues. A more comprehensive examination, utilising multiple software packages or multiple sets of parameters, would yield a larger set of residues, more accurately reflecting the set whose variation could impact antibody binding.

Analysis of the Conservation of Each Epitopic Residue

The purpose of the conservation analysis employed in these studies is to determine the likelihood of antigenic escape by assessing the degree to which each residue in the epitope has varied in the past, as evidenced from available amino acid sequences. As we saw in the above section, one problem with this approach is that the epitope is defined narrowly, so that some atomic interactions are likely to be missed. A further problem is that escape can be influenced by residues that are not part of the epitope *per se*. For example, an escape mutant raised against antibody CR6261 in an H5 strain exhibited the single substitution H111L in HA2. Residue 111 is buried in the pre-fusion state and is unlikely to be identified as an epitopic residue in any analysis, but the substitution is believed to cause a reorientation of the antibody binding pocket (Throsby et al., 2008). For these reasons, the studies may overlook important variation in other nearby residues.

Attention is often focussed on the 'highly conserved' A chain of HA2. Ekiert et al. (2009) assert high conservation of 11 residues on the chain (41, 42, 45, 46, 48, 49, 52, 53, 55, 56, 57). In 8 residues, the conservation quoted is > 99%, based on an analysis of available sequences across all subtypes. This analysis, and a similar analysis of residues 30-36 (Ekiert et al., 2011), is based on groupings of similar amino acids, with transitions within a group being classed as conserved. The groups used are: 1) Asp/Asn/Glu/Gln; 2) Phe/Tyr; 3) Ile/Leu/Val/Met; 4) Lys/Arg; 5) Ser/Thr (Ekiert et al., 2011).

The antibody/protein interface is subtle, and the precise nature of the atomic interactions involved can impact the conservation of substitutions. In particular, both hydrophobic and polar interactions can play a part (Xia et al., 2012). A simplistic division into groups is questionable. The definition of the first group above (group 1) has specific difficulties in the light of substitution studies and known escape substitutions. Ekiert et al., in the same studies, report Asp19Asn in H3 and H7 strains leading to a 30-fold increase in CR9114 K_d, and Glu15Gln leading to a tenfold reduction in affinity. The following 'group-conserved' escape mutants in influenza glycoproteins are reported by other researchers: Ile244Met (Yewdell, Caton, and Gerhard, 1986); Asp190Asn (Tsibane et al., 2012); Asp127Glu (Schmeisser et al., 2012). In addition, Wang et al. (2011) report Ile89Leu as an escape mutant, but it is not clear whether this residue is part of the epitope, or induces a conformational change elsewhere that induces escape. Returning to the 'highly conserved' A chain of Ekiert et al. (2009), if we remove group 1 (Asp/Asn/Glu/Gln) given the 5 counter-examples cited above, 4 of the 11 residues become variable rather than conserved. HA2 46, for example, which is reported as 99.6% conserved in the study, has undergone two frequency switches in H3 strains: D46N in 1995/96 and a reversion N46D in 2005/07.

Follow-up assays of wild-type amino acid variants revealed by the conservation analysis

The preceding points will serve to highlight key potential deficiencies in such follow-up assays: namely that they are unlikely to address all amino acid locations that could be involved in interactions, and that questionable assumptions of conservative similarity may lead to some substitutions being overlooked. Dreyfus et al. (2012) provide a particularly thorough analysis, covering nearly all substitutions observed in the wild in all identified epitopic residues of antibody CR9114. Interestingly, some substitutions outside the deduced epitope are considered, some of which (such as Q34T and Q34R) are seen to have a significant impact on K_d – findings that would appear to cast doubt on the completeness of epitope discovery. However, even this extensive analysis, covering 58 isolates, features only a small subset of the combinatorial possibilities. Such analyses can provide valuable insight, but they cannot provide a comprehensive understanding of the likelihood of escape.

Culture of escape mutants

Researchers have attempted to breed escape mutants in seven of the studies considered, and such mutants have been successfully bred in four (Table 5.6). It is interesting to note that Okuno et al. attempted to breed mutants from 10 base strains, and were only successful with 2.

Antibody	**Escape Mutant bred successfully?**	**Base Strain(s)**
C179	Yes	A/Suita/1/89 (H1), A/Izumi/5/65 (H2) (8 other strains were tried unsuccessfully)
CR6261	Yes	RG-A/Indonesia/5/05 (H5)
12D1 39A4	No	A/Hong Kong/1/1968 (H3)
A06	No	Not stated
D8 F10 A66	No	A/Vietnam/1203/04 (H5)
PN-SIA28	Yes.	A/PR/8/34 (H1)
CR8020	Yes.	A/Hong Kong/1/1968 (H3)
FI6	n.d.	
CR9114	n.d.	

Table 5.6: Escape Mutants against broad-spectrum antibodies, from the studies considered in this section. n.d.= not done.

It is therefore possible that escape mutants to antibodies with negative results in the studies could be bred from a different base strain to those employed.

Conclusion

While the affinity spectrum of broad-binding antibodies to the HA stalk is impressive, the experimental and analytical evidence reviewed in this section demonstrates that their action is not universal. Furthermore, when considering strains beyond those directly assayed, researchers have made simplifying assumptions relating to epitope coverage and amino acid similarity which are likely to over-state the binding breadth of the antibodies in question and to under-state the extent to which evolutionary variation in the stalk is possible. Even where neutralisation can be demonstrated *in vitro*, in over half the subtypes considered the concentration present *in vivo* is sufficiently low, on the basis of current understanding, as to cast doubts on the neutralising ability of the antibody. Returning to the question at the start of this section, it is clear from this analysis that there is scope even for the broad binding antibodies to present evolutionary pressure that the virus could reasonably be expected to overcome by evolution. Indeed, this is evidenced by the culture *in vitro* of escape mutants against such antibodies in four out of seven studies in which it was attempted.

5.6 Characterisation of the H3 Stalk Antigenic Regions

I will focus now on H3 because of the much larger number of HA2 sequences available compared to other subtypes. I first develop a list of residue locations comprising the antigenic sites in the stalk, following a similar process to that previously used to develop lists for sites A to E (Bush et al., 1999). That analysis built on earlier descriptions derived from substitutions in escape mutants (Wiley, Wilson, and Skehel, 1981; Wiley and Skehel, 1987; Wilson and Cox, 1990). Bush et al. used a set of 357 HA1 sequences to identify residues undergoing variation in the period 1968 to 1999. Where such residues were located directly in the area of an antigenic site and were not buried, they assigned them to the antigenic site, thus providing a broader and more comprehensive understanding of the delineation of the sites than could be obtained from escape substitutions alone. From a nucleotide-level analysis of the 357 sequences, 18 amino acid locations were found to have been under positive selective pressure during the period. All 18 are located within the identified antigenic sites, providing some confidence in the approach used to assign residues to sites. It is important to note that in the characterisation by Bush et al. only residues seen to vary between the strains under study were considered (this was clearly sufficient for the retrospective study in which the results were employed). Subsequent to the study period, variations occurred at many other locations that would otherwise have met the

criteria for assignment (Lees, Moss, and Shepherd, 2010). In characterising the stalk antigenic regions, I will consider both varying and to-date invariant locations.

The characterisation of sites in H3 HA2 is made easier by the availability of the three crystal structures of stalk-binding broadly neutralising antibodies bound to H3 HA2 (PDB codes 3SDY, 3ZTJ, 4FQY). These allow us to infer a more complete list of participating residues than would be available from the study of escape mutants alone, but the caveats noted in the previous section relating to epitope identification apply here also. To obtain a comprehensive view of the epitopes in these three structures, I took both the epitopic residues inferred by PDBsum from the three PDB structures (Table 3.1) and 13 additional residues listed by authors of the three studies, to create the combined view shown in Figure 5.4A. From this view, and from Figure 5.4B, it can be seen that antibodies FI6 and CR9114 (PDB codes 3ZTJ and 4FQY) bind in a similar location, while CR8020 (PDB reference 3SDY) binds in a separate location, closer to the viral membrane, with just four residues being shared across all three epitopes. I therefore mapped the locations listed in the figure into two antigenic sites: site F, embodying the epitope of FI6 and CR9114, and site G, embodying the epitope of CR8020 (Figure 5.4C). The overlapping locations were assigned to the closest of the two sites. Four residues were classified as 'outliers' and not assigned to either site F or site G. Two of these, HA1 277 and 278, lie in site C, while the other two, HA2 7 and 11, lie in an adjacent monomer (see Figure 3.1A).

A

Domain	Residue	3SDY	3ZTJ	4FQY	Assignment
HA1	19				G
	38				F
	40				F
	277				Outlier
	278				Outlier
	291				F
	292				F
	293				F
	318				F
	325				G
HA2	7				Outlier
	11				Outlier
	15				G
	16				G
	17				G
	18				G
	19				G
	20				F
	21				F
	25				G
	30				G
	31				G
	32				G
	33				G
	34				G
	35				G
	36				G
	37				F
	38				F
	39				F
	41				F
	42				F
	43				F
	45				F
	46				F
	48				F
	49				F
	52				F
	53				F
	56				F
	57				F
	146				G
	150				G
	153				G

B C

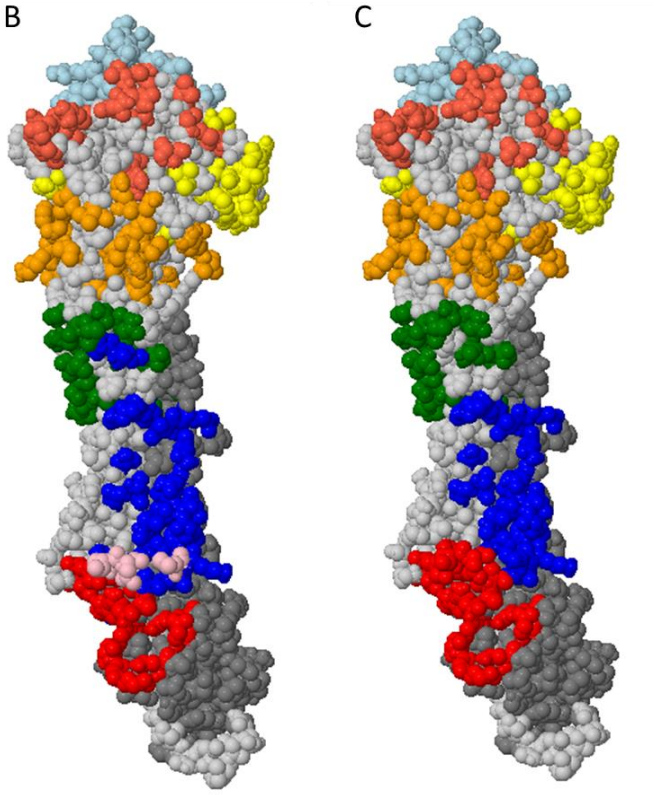


Figure 5.4: Epitopes deduced from three crystal structures of antibodies binding to the H3 stalk (PDB codes 3SDY, 3ZTJ and 4FQY). (A) Summary: Locations coloured black are inferred by PDBsum. Locations coloured grey are additional locations described by authors of the crystal studies. Residue locations highlighted yellow are those exhibiting variation amongst the H3 sequences in the database. The locations are assigned to stalk antigenic sites F and G on the basis of their location. (B) Epitopes of the three crystal structures shown on a single HA monomer: combined epitope of 3ZTJ and 4FQY in dark blue, epitope of 3SDY in dark red, common residues in pink. Canonical sites A-E are also shown: A-pale red, B-pale blue, C-green, D-yellow, E-orange. Other residues of HA1 are in pale grey, HA2 in dark grey. (C) Site F (blue) and site G (red), as described in the text and as assigned in 5.4A. Other colours as (B).

The majority of contacts in site F are in Helix A of HA2 (Figure 5.5A). HA1 and HA2 loops provide additional contacts. Antibody binding to Helix A (across all subtypes) is often, although not exclusively, associated with descent from the V_H1-69 germline (see Section 5.4). CR9114 is a descendent of this germline, while FI6 is not.

Site G is composed primarily of contacts in the outermost strand of a beta sheet at the base of HA2; contacts in a nearby short alpha helix (Figure 5.5B), and contacts in the fusion peptide (HA2 15-19). While the precise neutralising action of stalk-binding antibodies is not well understood, and more than one mechanism may contribute (Corti et al., 2011), binding to this latter set of contacts is likely to be important. Note that contacts to HA2 18 and 19 are common to all three antibodies.

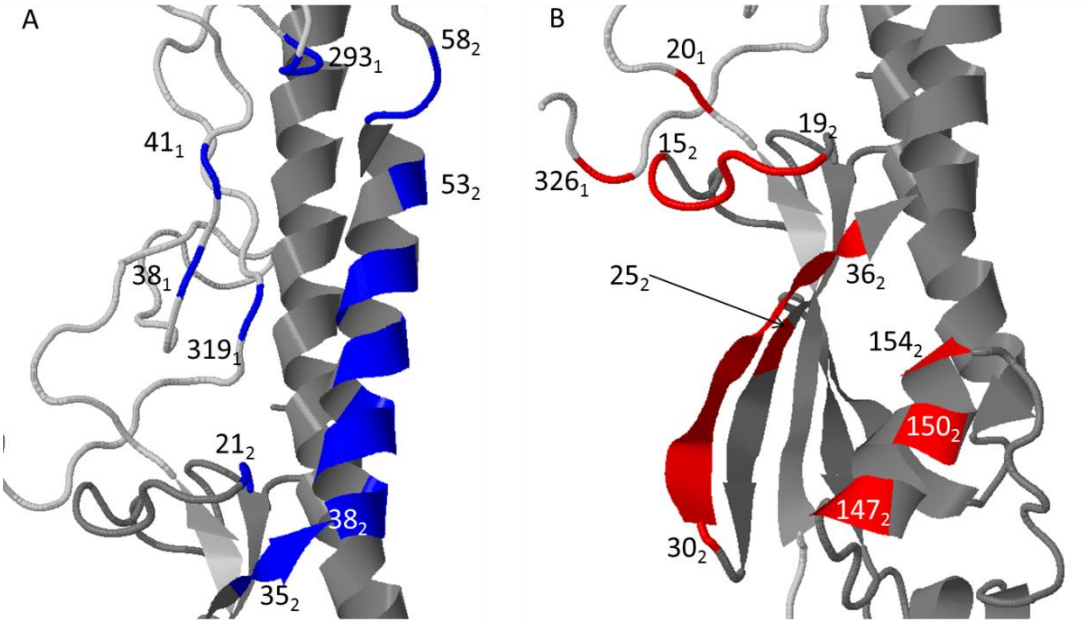


Figure 5.5: Detailed structure of site F (A, blue) and site G (B, red). HA1 in light grey, HA2 in dark grey.

5.7 Detection of Evolutionary Change in the Stalk

Recent statistical methods have extended standard approaches based on the ratio of synonymous to non-synonymous mutations (Bush et al., 1999) to take into account directional evolution towards "target" residues (Kosakovsky Pond et al., 2006) and the spatial clustering of mutations (Tusche, Steinbrück, and McHardy, 2012). However, all these approaches lack discriminatory power in regions such as the stalk where mutation rates are low: in Appendix A, I describe an application of the methods of Kosakovsky Pond et al. to the H3 stalk, and examine its power in further detail.

For this study I develop an approach for assessing the true degree of mutability in the stalk region, and draw conclusions relevant to broad spectrum antibody and vaccine development. In Chapter 3, I presented a method for predicting B-cell epitopes on the surface of HA1 that involved the identification of clusters consistent with the known dimensions of known B-cell epitopes; many clusters in regions of HA that mutate rapidly (notably the canonical antigenic sites) were detected between successive dominant seasonal strains, whereas clusters in other regions were rarely detected and only when mutations accumulated in dominant strains separated by much longer periods of time. Here I build on this approach in ways that are specifically designed to maximize the chances of detecting antigenic pressure in the H3 stalk region, where the overall mutation rate is generally low. Whereas the previous study considered simply the substitutions and fixations occurring between the dominant strains of the population, here I consider additional categories of mutation event occurring within the population, looking

for fixations and other variations that are co-ordinated both in distance and time, and compare the frequency of such events within the identified antigenic sites F and G, and outside them.

Viewed in isolation, mutations that never become universal in the global population provide relatively weak evidence of evolutionary pressure. However, multiple mutation events (both fixations and polymorphic events), occurring within close spatial and temporal proximity, provide strong evidence of selective pressure rather than hitchhiking. On the grounds that HA1 fixations are observed within the antigenic sites at a much greater frequency than outside them, hitchhiking is thought to be rare in HA1 (Shih et al., 2007). I make a similar argument here: given the low rate of fixation in the stalk, it is unlikely that fixations occurring closely together in space and time are caused by hitchhikers, which would be expected to be much more evenly distributed across the protein. In the following section, we will consider the non-uniform distribution of fixations in the stalk, which provides further substance to this argument.

The question remains as to whether co-ordinated events in the stalk can be attributed to antigenic pressure as opposed to adaptive mutations. Within the fusogenic region, adaptive mutations are known to change the pH of activation (Daniels et al., 1985; Thoennes et al., 2008; Reed et al., 2009; Reed et al., 2010), and such changes have been associated with the adaption of H5N1 strains to transmission in mammals, and the associated structurally substantial changes to receptor binding (Imai et al., 2012; Herfst et al., 2012). Adaptive changes to other aspects of membrane kinetics may also occur. It is therefore not possible to state with certainty whether a particular mutation is related to antigenic escape. In this analysis I present evidence of temporally co-ordinated mutations at multiple locations within epitope-sized patches in a region that is known to bind antibodies, which provides a strong overall case for antigenic pressure being one of the factors at work. Supporting evidence comes from Wang et al. (2011), who identify a mutation in this region in wild-type H1N1 strains leading to a reduction in the effectiveness of cross-neutralising antisera.

Daniels et al. (1985) bred viral mutants of the H3N2 X-31 strain in conditions of raised endosomal pH. Across the mutants sampled, they identified mutations at the following 13 locations: HA1 17, 102, 207 and HA2 6, 9, 47, 54, 57, 81, 105, 112, 114, 163. Of these, all but three are invariant in H3 samples, the three showing variation being HA1 207 and HA2 57, 114. If compensatory changes to regulate activation pH were a major factor in H3, we would expect to see more of the locations identified in the X-31 experiments mutating in wild-type samples. Of the 13 residues identified in the experiments, only HA2 57 lies within the stalk-based antigenic sites (in site F).

5.8 Fixations and Polymorphism in the HA Stalk

Both unvaried and mutable locations are found in all regions of HA. One common pattern of mutability is an abrupt fixation, in which the consensus residue at a location changes rapidly at a particular point in time (Figure 5.6A). This pattern, which above the stalk is almost exclusively found in the antigenic sites, has previously been associated with positive selective pressure, as discussed above and in Shih et al. (2007). Another pattern is that of transient polymorphism (Figure 5.6B). The punctuated nature of these events is again suggestive of selective pressure, although in this case, as the mutation does not become fixed in the population, it is likely that a separate mutation, providing greater fitness, emerges to take its place, or that the evolutionary pressure is either transitory or balanced out by other factors. In this analysis, in order to differentiate strong and weaker signals, the dominant amino acid must be present in less than 80% of samples in a year for a period to be classified as one of transient polymorphism, other periods of polymorphism that do not meet this threshold are classified as 'intermediate' (Figure 5.6C). Finally, any location at which the dominant amino acid is identical in all years and present in at least 95% of samples is classified as 'invariant'.

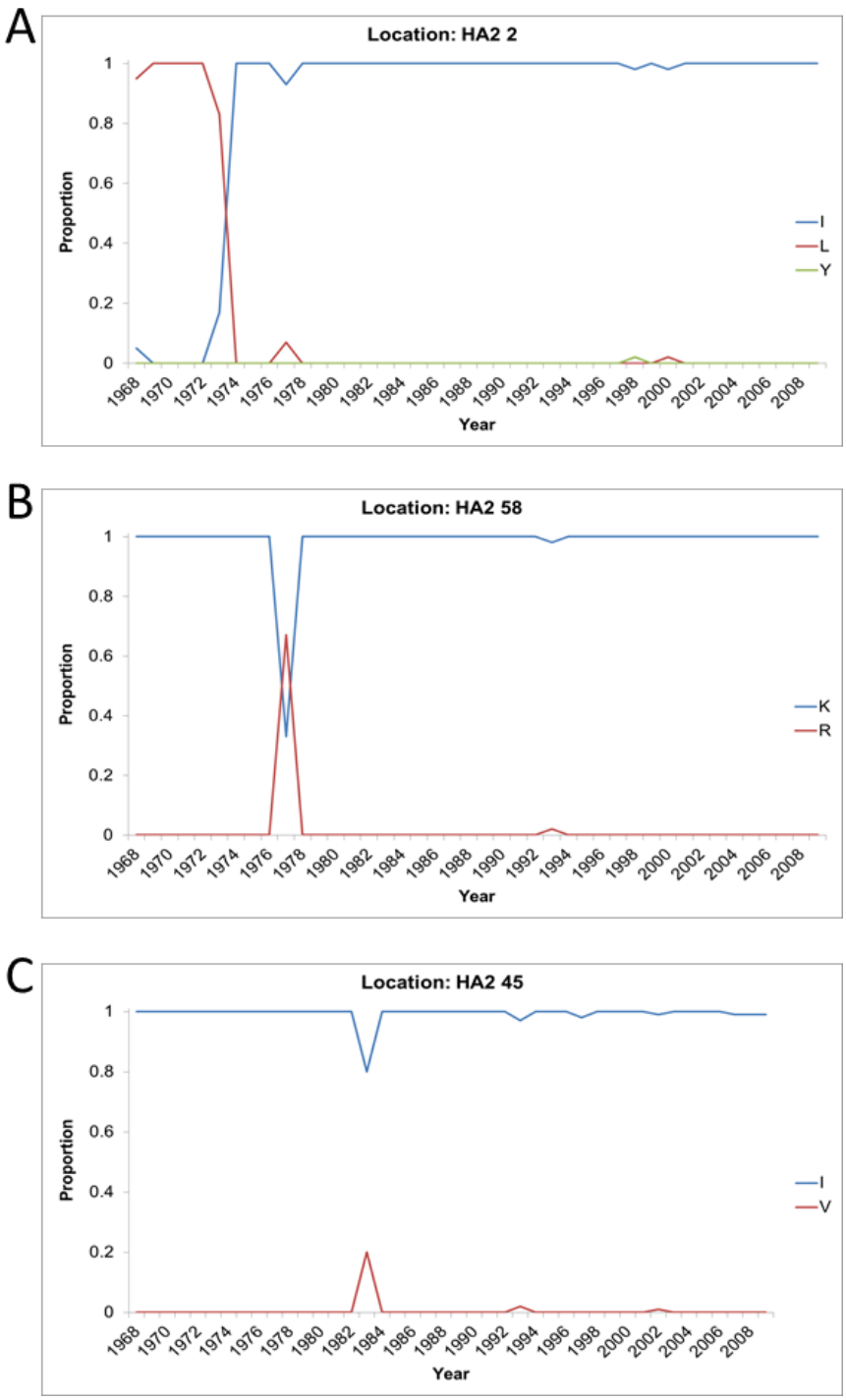


Figure 5.6: Patterns of mutability in selected HA locations: (A) – fixation, (B) – transient polymorphism, (C) – intermediate variation.

I classified each location in H3 HA based on observed activity across all years. Locations in which one or more fixations occurred over the period 1968-2009 were classed as 'Switchers', while locations exhibiting transient polymorphism but no fixations were classed as 'Polymorphic'. Likewise further sites were classified as 'Intermediate' or 'Invariant' (Table 5.7).

Classification	HA1 Residues	HA2 Residues
Switcher	2, 3, 25, 31, 50, 53, 57, 62, 63, 75, 78, 82, 83, 94, 122, 124, 126, 131, 133, 135, 137, 140, 142, 143, 144, 145, 146, 155, 156, 157, 158, 159, 163, 172, 186, 188, 189, 190, 192, 193, 196, 197, 202, 207, 213, 217, 222, 225, 227, 242, 244, 260, 262, 275, 276, 278, 299	2, **32**, **46**, 55, **57**, 121, **150**
Polymorphic	5, 8, 9, 10, 16, 33, 34, 47, 54, 67, 81, 88, 92, 95, 96, 98, 106, 112, 121, 138, 160, 164, 167, 173, 174, 182, 185, 199, 201, 209, 216, 219, 226, 230, 247, 248, 257, 271, 273, 280, 294, 304, 307, 323	**18**, **56**, 58, 66, 97, 102, 110, 123, 124, 126, 150, 158
Intermediate	4, 6, 7, 13, **19**, 23, 27, 44, 45, 48, 49, 55, 79, 80, 103, 105, 118, 120, 128, 129, 132, 141, 162, 169, 171, 175, 176, 179, 187, 194, 198, 208, 212, 214, 218, 220, 223, 229, 232, 233, 236, 240, 245, 246, 252, 261, 264, 265, 267, 268, 269, 279, 282, **291**, **292**, 297, 298, 303, 308, 309, 310, 312, 322, **325**, 326, 327	4, 11, 27, **43**, 44, **45**, **49**, 51, 59, 71, 79, 82, 84, 98, 100, 103, 106, 113, 114, 120, 139, 143, 155, 156, 171
Invariant	1, 11, 12, 14, 15, 17, 18, 20, 21, 22, 24, 26, 28, 29, 30, 32, 35, 36, 37, **38**, 39, **40**, 41, 42, 43, 46, 51, 52, 56, 58, 59, 60, 61, 64, 65, 66, 68, 69, 70, 71, 72, 73, 74, 76, 77, 84, 85, 86, 87, 89, 90, 91, 93, 97, 99, 100, 101, 102, 104, 107, 108, 109, 110, 111, 113, 114, 115, 116, 117, 119, 123, 125, 127, 130, 134, 136, 139, 147, 148, 149, 150, 151, 152, 153, 154, 161, 165, 166, 168, 170, 177, 178, 180, 181, 183, 184, 191, 195, 200, 203, 204, 205, 206, 210, 211, 215, 221, 224, 228, 231, 234, 235, 237, 238, 239, 241, 243, 249, 250, 251, 253, 254, 255, 256, 258, 259, 263, 266, 270, 272, 274, **277**, 281, 283, 284, 285, 286, 287, 288, 289, 290, **293**, 295, 296, 300, 301, 302, 305, 306, 311, 313, 314, 315, 316, 317, **318**, 319, 320, 321, 324, 328	1, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, **15**, **16**, **17**, **19**, **20**, **21**, 22, 23, 24, **25**, 26, 28, 29, **30**, **31**, **33**, **34**, **35**, **36**, **37**, **38**, **39**, 40, **41**, **42**, 47, **48**, 50, **52**, **53**, 54, 60, 61, 62, 63, 64, 65, 67, 68, 69, 70, 72, 73, 74, 75, 76, 77, 78, 80, 81, 83, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 99, 101, 104, 105, 107, 108, 109, 111, 112, 115, 116, 117, 118, 119, 122, 125, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 140, 141, 142, 144, 145, **146**, 147, 148, 149, 151, 152, **153**, 154, 157, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 172, 173, 174, 175

Table 5.7: H3 HA locations classified by mutability. Locations in antigenic sites F are shown in bold blue and those in site G in bold red.

Figure 5.7A illustrates the distribution of these classifications across the H3 HA structure, and in Figure 5.7B I show a similar analysis of H1 locations for comparison, although it should be noted that there is only limited availability of H1 sequences for years before 2000, meaning that there is a greater likelihood of classification error, probably accounting for the larger number of switcher and polymorphic locations.

Invariant residues are found in all regions of the molecule. In H3 HA1 they delineate the antigenic sites. In H3 HA2, while the overall level of mutability is reduced compared to that in HA1, mutable locations do occur, and are grouped in specific locations. If mutations in HA2 were purely hitchhikers, one would expect a random distribution. This grouping is therefore indicative of selective pressure. While, in other areas of H2, some invariance is mandated by structural or functional considerations, in other cases it may be attributable to weak (or absent) selective pressure.

Fixations have occurred at 7 locations in HA2, of which 4 are located in the antigenic sites F and G. If fixations were distributed evenly in HA2, such a concentration in these sites is unlikely ($p < 0.03$[2]). Two of the remaining three fixations, HA2 55 and HA2 121, occur close to the sites. HA2 55 is immediately adjacent to HA2 56, which is located in site F. It is a 'buried' residue, but, as evidenced by the escape mutation raised against CR6261 by Throsby et al. (2008), buried residues can elicit antigenic escape. HA2 121 lies near to the boundary of both antigenic sites and is surface-exposed in the trimeric structure (Figure 5.8). Its Cα atom lies at 10Å from HA2 153, the nearest residue in site G, and 12Å from HA2 39, the closest residue in site F. It is sufficiently close to participate in the epitope of antibodies binding in either site F or site G, although its orientation on the three-dimensional molecule makes the former more likely than the latter as it is roughly co-planar with regions of site F, but lies over a ridge from site G.

[2] As there are 175 locations in HA2 and 34 of them are in the sites F and G, the probability of four fixations lying in sites F and G, and three fixations lying outside the sites, assuming a random distribution, is

$$\frac{(34 \times 33 \times 32 \times 31 \times 141 \times 140 \times 139)}{(175 \times 174 \times 173 \times 172 \times 171 \times 170 \times 169)} \times {}_7C^4 = 0.024$$

and the probability of four or more fixations lying in the sites is 0.027.

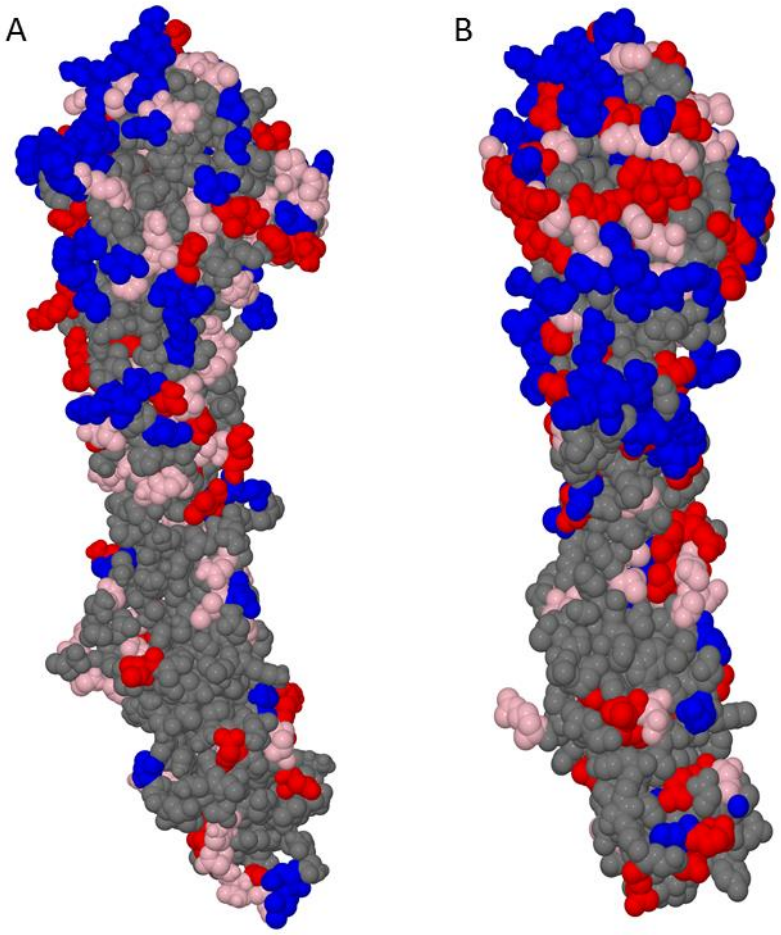


Figure 5.7: Locations in H3 HA (A) and H1 HA (B) classified by mutability. Blue – switcher; red – polymorphic; pink – intermediate; grey – invariant.

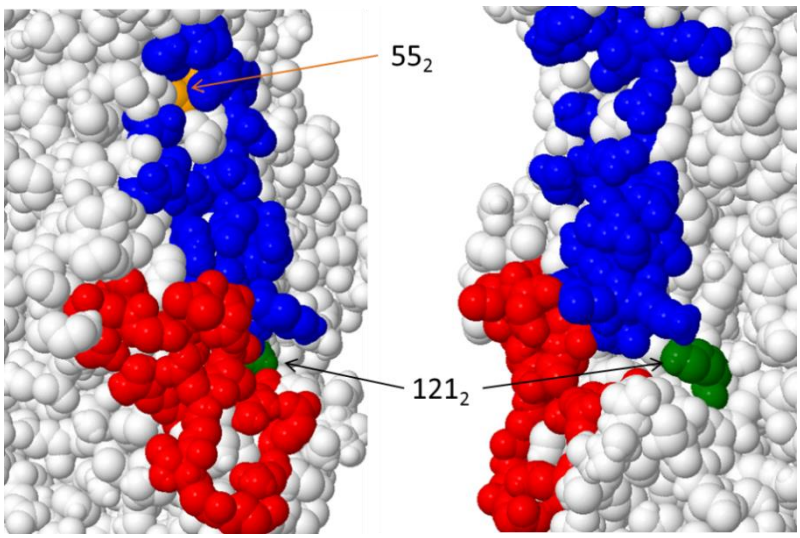


Figure 5.8: Fixations occur at HA2 locations 55 (brown) and 121(green), which lie outside but adjacent to the characterised antigenic sites F and G. The residues are shown in two views of the HA monomer. Site F is shown in blue and site G in red: other residues in white.

N-glycosylation sites are known to interfere with the attachment of antibodies (Skehel et al., 1984). They are characterised by the motif Asn-X-Ser/Thr (where X is not Pro), but the likelihood of such a motif being populated with a glycan is dependent on the specific sequence and its context. The server NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc), which employs artificial neural networks, is widely used to predict populated sites, and is claimed to have 76% accuracy when tested against known populated and non-populated sites (Gupta, Jung, and Brunak, 2004). When the server is run against the full-length HA sequences of A/Hong Kong/1/1968 and A/Perth/16/2009, with a potential in excess of 0.5, which is the default threshold, it predicts the same four sites in the stalk of each protein. The asparagine residues of the sites are at HA1 8, 22, 38 and HA2 154 (Figure 5.9). The residues HA1 8-10 display some polymorphism between 1971-1972 and 1980-1982. Other constituent residues of these N-glycosylation sites are invariant, apart from some slight mutability in HA2 155 in 2004-2005. The four sites are therefore likely to be permanent in H3, apart from some exceptions during the polymorphic years of HA1 8-10. Experimental evidence confirms that all four sites are occupied in A/Hong Kong/1/1968 and A/Aichi/1/1968 (Wilson, Skehel, and Wiley, 1981; Gallagher et al., 1988).

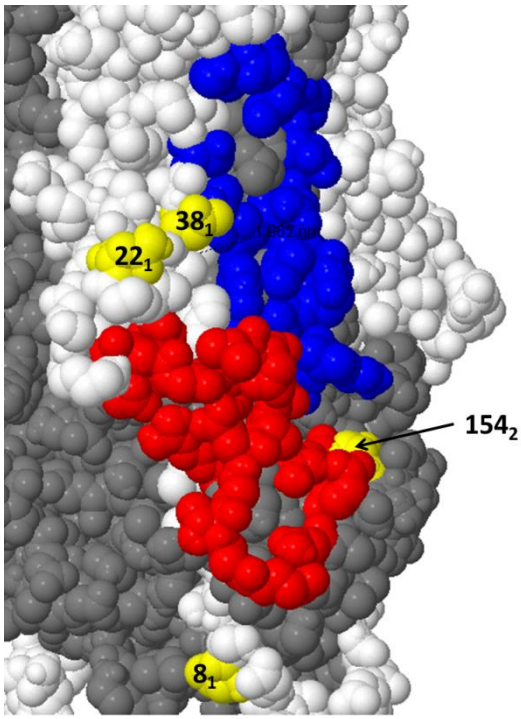


Figure 5.9: N-glycosylation sites of the HA H3 stalk, predicted by the NetNGlyc server from full-length sequences of A/Hong Kong/1/1968 and A/Perth/16/2009, shown on the HA trimer. Asparagine residues of the glycosylation sites are shown in yellow. Antigenic site F is in blue and site G in red. Other HA1 residues are white, HA2 grey.

Despite the known ability of N-glycosylation sites to interfere with antibody binding, HA1 38 and 39 are included in antigenic site F, and HA2 154 is immediately adjacent to site G. It is

unlikely that an antibody could bind across a populated N-glycosylation site (this was the specific point established experimentally by Skehel et al.). If we assume that neutralising antibodies in HA2 must bind to the fusion peptide, the two N-glycosylation sites with asparagines at HA1 22 and 38 would serve to prevent neutralising antibody binding extending to the left of site F or above the top of site G, thus explaining the demarcation of the sites. This in turn could explain the lack of mutability in HA2 in the region beyond the two N-glycosylation sites: while it may be possible for antibodies to bind there, their binding would be unable to extend across the sites to reach the fusion peptide. They would therefore not have a neutralising action and so would not exert evolutionary pressure.

5.9 Antigenic Transition in Antigenic Sites F and G

In this section, we will examine mutability in antigenic sites F and G, and in the two neighbouring switcher locations. We will see that, at various points since the introduction of H3 to the human population in 1968, change has occurred across multiple locations in the sites, in a manner that is co-ordinated in time, and that the affected locations form clusters whose dimensions are consistent with the dimensions and spatial constraints of B-cell epitopes.

Figure 5.10 shows a timeline of mutable events in the antigenic sites and the two neighbouring switcher locations. We shall consider these events in chronological order, considering the possibility of linkage between events that are both spatially and temporally close.

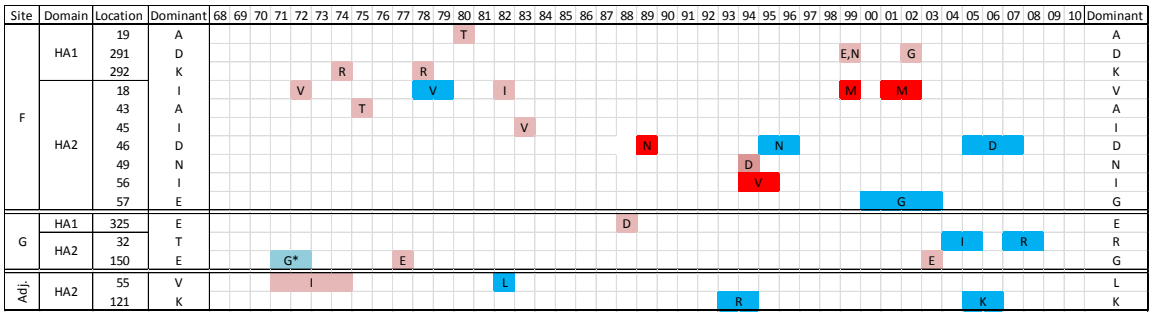


Figure 5.10: Year-by-year variation in mutable residues of antigenic sites F and G, and in the nearby locations HA2 55 and 121 where fixations occur. The dominant residue in 1968 and in 2010 is shown in the 'Dominant' columns at each end of the table. Fixations are shown in blue, together with the code of the newly fixed residue. Periods of transient polymorphism (where the frequency of the dominant residue falls below 80%) are shown in dark red. Other periods of polymorphism (where the frequency of the dominant residue falls below 95% but not as low as 80%) are shown in pink. In cases of polymorphism, the code of the minority residue is shown. Because of the low number of HA2 sequences in early years, the frequency switch in HA2 150 in 1971/72 identified from available sequences does not pass the test for statistical significance (Shih et al., 2007). We can be confident that the dominant residue was E (Glutamic Acid) in 1968 samples, as A/Hong Kong/1/1968 is well documented, and G (Glycine) from 1972 onwards, as there are sequences available for 10 strains that year and all have 150G, but the exact point of transition cannot be reliably determined as there are only 4 HA2 sequences available for each year from 1969 to 1971.

In the period 1971-1983, a fixation in site F is co-ordinated with four polymorphisms in the same site. A period of polymorphism in the adjacent location HA2 55 occurs between 1971 and 1974 during which the minority residue is isoleucine, and the period ends with a fixation in that location to leucine. A fixation also occurs at HA2 150, but this location is towards the bottom of site G and therefore unlikely to be part of a linked antigenic cluster. Excluding HA2 150, the cluster distance of the other 6 residues is 30Å, which is consistent with the dimension established in Chapter 3 for a B-cell epitope (Figure 5.11A). The temporal co-ordination between changes at locations HA2 18 and HA2 55 is striking (Figures 5.11B, C). There is polymorphism at both locations in 1972. Following the fixation in HA2 I18V in 1978, the fixation HA2 55 in 1982 is accompanied by a 25% reversion to isoleucine at HA2 18.

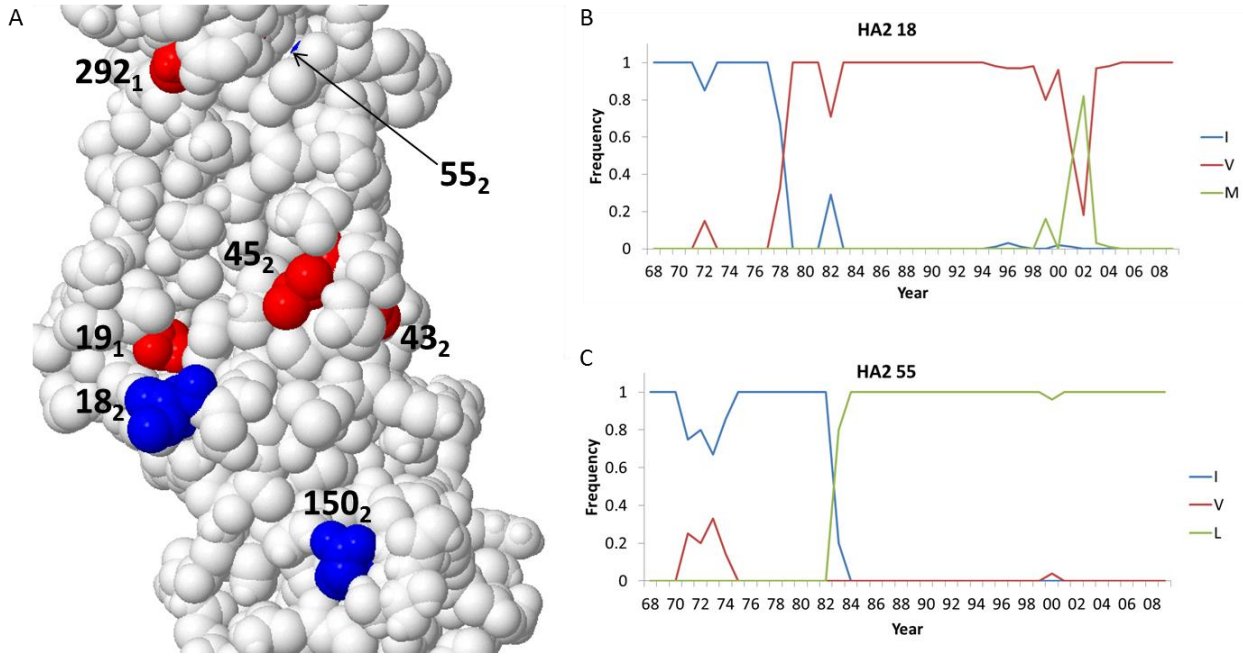


Figure 5.11: Locations displaying variation in the period 1971-83: (A) fixations in blue, others in red. (B),(C) Amino acid frequency charts for this period for locations HA2 18 and HA2 55 respectively.

Following 10 years in which little activity is seen, a fixation HA2 K121R in 1993 is followed by transient polymorphism at HA2 49 and 56 in 1993-1994, and then by the fixation HA2 D46N in 1996. Three of the affected locations are in antigenic site F, while the fourth is the neighbouring residue HA2 121. The locations are roughly collinear, with a cluster distance of 34Å, which is again consistent with the dimensions of a B-cell epitope (Figure 5.12).

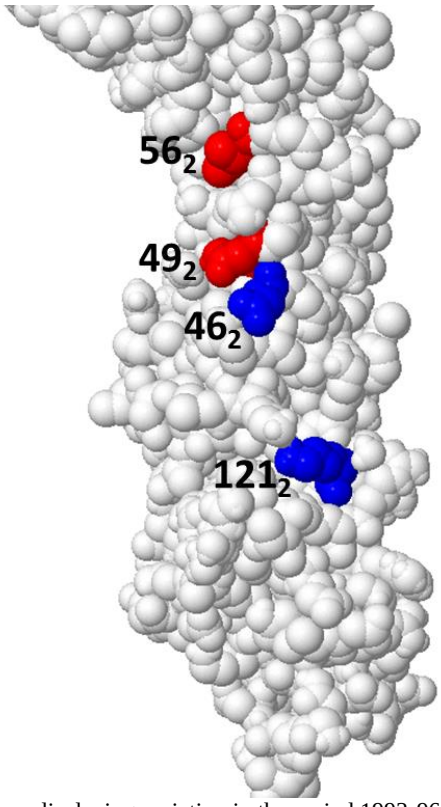


Figure 5.12: Locations displaying variation in the period 1993-96: fixations in blue, others in red.

In 1999, polymorphic activity is seen in site F at HA1 291 and HA2 18 (Figure 5.13). This is followed in 2000 by a fixation in HA2 57, the period of fixation extending over four years, during which time there is continued polymorphism at the first two locations. The activity in site F concludes with the fixation HA2 N46D: this is a reversion of the fixation observed in the previous period.

Coincident activity in site G starts with a small transient polymorphism in HA2 150 in 2003, following which there are two fixations at HA2 32: T32I in 2004/5 and I32R in 2007/8. These changes are quite distant from the changes in site F – the overall cluster size is 45Å, and so (on the assumption that the changes are indeed related to antigenic escape) are likely to be independent.

Also on this period, there is a fixation neighbouring the antigenic sites: HA2 R121K. This is a reversion of the second fixation observed in the previous period. As noted in the previous section, HA2 121 is close to both sites F and G, but a ridge separates it from site G, so that it is more likely to be part of an epitope with antibodies binding in site F. It also seems most likely that the two reversions occurring in this period are related rather than independent, particularly as the two fixations are closely associated.

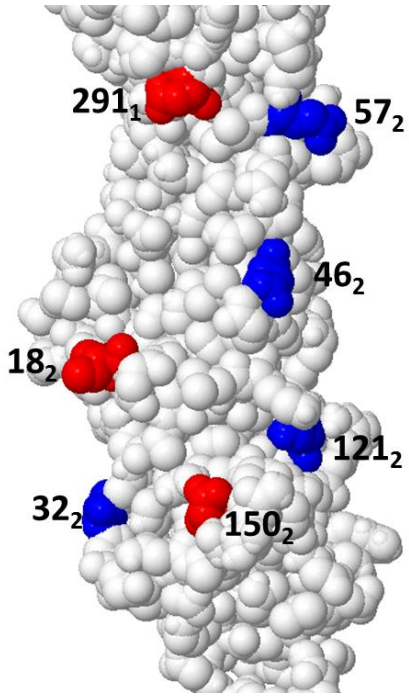


Figure 5.13: Locations displaying variation in the period 1999-2008: fixations in blue, others in red.

The reversions at HA2 46 and 121 suggest a relaxation of the evolutionary pressure under which the substitutions developed, but, if we assume that this pressure was imposed by a neutralising antibody, one would expect the antibody to persist. One alternative scenario is as follows. In the period 1993-96, antigenic pressure from an antibody binding in the site induced two substitutions (in HA2 46 and 121), leading to escape but reducing fitness. In the period 1999-2008, pressure from a further antibody induced further substitutions, in HA2 46 and 57. If these new substitutions also provide escape from the first antibody, the substitutions at HA2 46 and 121 can be reversed, increasing overall fitness. Another possibility is simply that the substitution at HA2 57 emerged as a fitter alternative to the previous two substitutions, which still provided for antigenic escape.

5.10 Discussion

In this study I have presented evidence of directed evolution in the H3 stalk. Crucially, this evidence draws on factors that are not considered by existing statistical approaches, extending the analysis of Shih et al. (2007) to HA2 and extending it to include consideration of the temporal coordination of fixations and polymorphisms and their locations with respect to known epitopes.

Some authors have expressed hope that the elicitation of broadly-based anti-stalk antibodies can provide universal immunity against influenza (Russell, 2011; Nabel and Fauci, 2010), however those antibodies discovered to date may be more restricted in their binding ability at

physiological concentrations. In addition, the assumption that a 'conserved' substitution will not impact antibody binding has led to unwarranted assumptions concerning the stalk's level of conservation. That HA escape mutants to broadly binding antibodies have been bred successfully in four cases also sounds a note of concern. I have reviewed evidence from a number of studies that neutralising stalk binding antibodies – albeit not so broadly-binding – are widely present in the human population. Taken together, these points suggest that neutralising stalk binding antibodies form a continuum of binding ranges and neutralising abilities. The broad-binding examples from the literature may lie towards one end of this continuum, but they are not essentially different. Extending vaccination techniques to elicit a greater quantity of stalk-binding antibodies is worthwhile, but it is not obvious that such antibodies will meet the very high expectations being placed on them in some quarters.

Within H3 strains, neutralising anti-stalk antibodies have been found to bind in two sites, which I have termed sites F and G. These sites overlap with the functionally active 'fusogenic region' of HA, and, while the exact neutralisation mechanism of such antibodies is uncertain, it is reasonable to assume that interference with this function is key to their neutralising capability. It is notable that mutable activity (in the form of fixations and polymorphisms) is significantly more frequent in sites F and G than in other regions of HA2. This provides strong evidence that such activity is directed and not simply the result of hitchhiking. We cannot with certainty ascribe any particular mutation or set of mutations to antigenic pressure – but, with the known presence of antigenic activity in the region, and the documented possibility of antigenic escape, there is a strong case for antigenic pressure to be considered as one factor. I have presented a possible explanation of the mutable activity observed in the H3 stalk that is consistent with antigenic pressure as a primary cause. This argument cannot be conclusive, but deserves further study.

These results have consequences for influenza surveillance, vaccine selection, and vaccine design.

It is important to monitor the development of polymorphism in the stalk, as it may be indicative of antigenic escape with implications for sections of the population that have developed protection via relatively broadly protective antibodies binding there. Supporting evidence comes from studies of pandemic H1N1 strains, where the mutation E374K in the stalk has been associated with vaccine breakouts (Maurer-Stroh et al., 2010; Strengell et al., 2011). As an associated point, serological surveillance has tended to focus on changes in HA1. It would be helpful to develop assays that will identify antigenic escape from stalk-binding antibodies, which could be indicative of strains to which the wider population has reduced immunity.

Assays based on artificial or augmented constructs (Martínez-Sobrido et al., 2010; Pica et al., 2012) may help in this respect.

It is unlikely that a single human antibody to the HA stalk could confer universal protection. Even if such an antibody could be isolated or developed, despite the points above on relative conservation and binding strength, there would still remain the difficulty of eliciting that precise antibody in all human phenotypes. There appear to be two approaches to developing a universal vaccine. One is to target a single epitope that, for functional reasons, is truly conserved across strains and subtypes of interest. The other is to develop antigenic 'defence in depth' in which antibody response is elicited to a number of relatively well conserved epitopes, such that escape from any one antibody will not of itself confer evolutionary advantage. On the basis of the analysis in this chapter, the HA stalk is not a good candidate for the first approach, but it has potential for the second. When eliciting antibodies to the HA stalk in a vaccine, it would be advisable to present multiple stalk antigens, representing the variation found in strains to date.

6 Conclusions and Areas for Further Study

6.1 Conclusions

A rich collection of influenza sequences, covering the past 45 years, is publicly available via GenBank (http://www.ncbi.nlm.nih.gov/GenBank) and specialist influenza resources. The body of knowledge continues to grow. In particular, the US National Institute of Allergy and Infectious Diseases has conducted, since 2005, a project to provide whole-genome sequences of early strains, for which, previously, only an HA1 sequence was available (National Institute of Allergy and Infectious Diseases, 2013). Where multiple sequences of the same strain exist in GenBank, correspondence is generally good. Exceptions exist for some early strains, where some erroneous sequences have been used in earlier studies (Section 2.4.1).

To my knowledge, no public database of antigenic data exists. The combination of sequence and antigenic data is valuable, as evidenced by this study, and I am exploring options to make the data publicly available to other researchers. The systematic collection of antigenic data has made it possible to examine the repeatability and accuracy of HI assay results. Where there are multiple measurements of the same strain/serum pair, the standard deviation is generally below one unit of antigenic distance, and the generally accepted view that assay results are accurate to within +/– 1 unit of distance is reasonable. 'Two-way' Archetti-Horsfall distances have greater precision (Tables 2.1-2.4).

I have demonstrated the use of substitution analysis in combination with an understanding of the three dimensional structure of HA to predict the location of those epitopes in wild-type human HA which have driven antigenic evolution. While experimental work would be needed to confirm the accuracy of these predictions, the good correspondence between 'immunoactive' locations in HA1 identified by this method and by other methods (Section 3.4) is encouraging. A simple predictive model of antigenic distance, based on changes at the immunoactive locations identified via this analysis, is capable of matching or exceeding the predictive powers of other models published to date, while encapsulating a greater degree of generality, which may lead to superior performance when addressing novel strains (Section 3.8). Such predictive models may prove of value where a rapid determination is required, where difficulties are encountered with conventional assays (Section 2.2), or where such a large-scale analysis is required that laboratory assays would be impractical (Section 6.4).

Results from the analysis support the view that, in general, antibodies to H3N2 HA1 bind in the five 'canonical' antigenic sites A-E. The survey of HI assay quality demonstrates that the

immune response across a number of individual animal hosts (in this case, ferrets) is, at a gross level, reasonably consistent. These points suggest that, firstly, epitope-bearing surfaces of proteins do have some distinguishing properties, and, secondly, that the antibody germline has a strong influence on locational response.

The epitopes predicted by substitution analysis were grouped in two key regions in H3 HA1 (Section 3.3). Through the use of an antigenic map simulation, I provided support for the hypotheses that either this property, or the tendency in some epitopes for the binding energy to be concentrated in a small number of residues, could lead to the clustering phenomenon seen in H3N2 antigenic maps and that the two could act in combination (Section 4.5). In the course of this work, I demonstrated the use of density-based clustering algorithms in antigenic maps, and introduced a Silhouette-based metric for determining the degree of cluster quality.

Although HA1 and HA2 overlap in the membrane-proximal region of the globular head, HA2 residues play only a small part in the epitopes predicted by substitution analysis (Section 5.3). A number of antibodies to the HA stalk have recently been characterised, and I map their binding locations to the H3 stalk to two new antigenic regions, F and G (Section 5.6). These broadly-binding antibodies may not be broadly neutralising at physiological concentrations, and, while the regions in which they bind are conserved in comparison to the HA head, substitutions observed in the wild may be sufficient to allow antigenic escape (Section 5.5). Their existence therefore does not rule out the possibility of antigenic evolution in regions F and G, and, while other evolutionary factors are likely to be at work in the region, the possibility of antigenic evolution should not be ruled out (Section 5.9). Such antigenic evolution would have significant consequences for the development of vaccines based on the elicitation of broadly-binding anti-stalk antibodies (Section 5.10).

6.2 Areas for Further Study

6.2.1 Epitope Prediction

The techniques used in this study identify clusters of substitutions that are plausible as epitopes in terms of their spatial and temporal grouping. It is unlikely that all locations identified within the identified clusters are epitopic. The technique could be refined further by checking the predicted epitope for structural plausibility, and by checking it against the known general properties of epitopes, such as planarity and linearity (Rubinstein et al., 2008; Kringelum et al., 2013). Some caution would be necessary, however: as discussed in Sections 3.2 and 5.4 and in the cited studies, antibody binding is opportunistic in its nature, and many binding

configurations are possible. It might be possible to develop a probabilistic approach, in which the properties of a predicted epitope are compared against a database of known structures in order to give a likelihood score. This approach would assume, though, that anti-HA antibodies conform to a general norm of epitopes in other proteins. It could be that the shape of the protein is atypical when compared to other antibody targets. Likewise, it may turn out that an antibody's germline descent has a significant bearing on the properties of its epitopes – as was suggested for the IGHV1-69 germline in Section 5.4, in which case a generic approach may not be meaningful.

Another enhancement that could add specificity would be to examine the fitness cost of a particular substitution, in the absence of antigenic activity, by measuring the change in free energy caused by the substitution, or by assessing the novelty of the substituted residue in that location. Again, though, some caution would be necessary: as noted in Section 5.5, even apparently conservative substitutions can elicit escape.

6.2.2 Simulated Mutagenesis

A wider study of epitope mutagenesis is needed in order to fully understand the scope of the identified broad-binding antibodies and the possibilities for escape (Section 5.5). The large number of HA/antibody structures now available, together with experimental mutagenesis studies to date, would provide a useful starting point for simulated mutagenesis studies using molecular dynamics (MD). While such studies are notoriously computationally expensive - one HA/antibody MD study was conducted on an IBM supercomputer (Zhou, Das, and Royyuru, 2008) – computational power continues to increase. Specialist hardware is available (Shaw et al., 2007), as are increasingly large clusters of conventional computers. The available experimental data could provide a useful platform for testing algorithms and optimisations, either generic or problem-specific.

6.2.3 dN/dS Studies

The study of nucleotide synonymous and non-synonymous mutation rates is a widely accepted approach for identifying locations or regions under positive or negative selective pressure; however, as discussed in Appendix A, as generally developed the approach lacks sensitivity when applied to the sequence data available for HA, and embodies some questionable assumptions. Refinements highlighted and suggested in the Appendix, in particular applying a dN/dS approach to a sliding window of isolation years, might address the sensitivity issue to some extent by giving greater prominence to fast-moving events such as rapid fixations.

6.2.4 Neuraminidase

Although antibodies binding to NA do not appear to impede viral cell entry, they select for escape mutants (Webster, Hinshaw, and Laver, 1982), indicating that they have some neutralising action. It has been suggested that including NA in vaccines provides enhanced protection, although evidence is mixed (Johansson and Brett, 2008; Nayak et al., 2010). There is support for the occurrence of antigenic drift in NA: the substitution rate is similar to HA, and a crystal study exists of an antibody which binds close to the active site, in a region known to have undergone significant variation in wild-type strains (Venkatramani et al., 2006; Air, 2011). Extension of the analysis to NA could provide some useful insight into its antigenic activity, for example in determining the extent of antigenic sites and the typical protection period. I have already started some work in this direction: the database contains a collection of N2 sequences, and a visualisation of sequence differences on the NA head is available.

6.2.5 Other Viruses

The techniques developed in this study are potentially extensible to other RNA-based viruses that exhibit antigenic drift, and for which there are sizable sequence data sets available. Hepatitis C virus (HCV) and HIV are possible candidates, each having well in excess of 100,000 sequences in GenBank (http://www.ncbi.nlm.nih.gov/GenBank/). In contrast to influenza, these two viruses cause persistent infection in humans. Study of sequences obtained from individual patients over time with these techniques might provide insights into antibody and viral development. Kwong and Wilson (2009) provide an interesting summary of antibody recognition of HIV and parallels with broad-spectrum influenza HA recognition. While the humoral response plays a clear role in HIV infection, its role in HCV infection is not well understood. Rapid viral clearance in some individuals is associated with the T-cell response. While B-cells are elicited, particularly in later infection, the majority of antibodies secreted are not HCV specific (Rehermann, 2009): HCV may therefore not prove to be a good candidate for discovery of antibody-induced evolutionary pressure.

6.3 Laboratory Investigations

A number of laboratory investigations, which would help to advance the overall aim of this work, are highlighted in this section.

6.3.1 Assays

The HI assay, widely used in surveillance, is only useful for detecting the neutralising action of antibodies that sterically block host cell binding. In recent years, the assay has suffered from problems: in particular the H3N2 subtype's gradual loss of affinity to avian red blood cells, and a receptor binding site in N2 neuraminidase (see Section 2.2.3). As a result, for H3 strains, WHO collaborating centres have started to introduce the microneutralisation assay alongside the HI assay. The WHO protocol for the microneutralisation assay is a two-step process (World Health Organization, 2011). In the first step, virus is mixed with dilutions of serum, and time is allowed for binding to take place. In the second step, MDCK cells are added, and time is allowed for viral infection. At the end of this second step, an ELISA assay is used to identify infected cells. The protocol is designed to assay the effectiveness of antibodies that act on the virus before cell entry. It is evidently less sensitive to those that act on budding – a mechanism that has been suggested both for stalk binding anti-HA antibodies (Corti et al., 2011) and for anti-NA antibodies (Air, 2011). A modified assay, which could be run at high throughput for surveillance purposes, but was equally sensitive to neutralisation at any point in the replication cycle, would facilitate the identification of such antibodies and provide a more complete picture of the neutralising potential of antisera. An early step in this direction was undertaken in collaboration between Birkbeck and the UK WHO CC at the suggestion of Dr John McCauley and under the supervision of Dr John McCauley and Dr Adrian Shepherd, but the results were inconclusive in the identification of such antibodies (Akere, 2012).

6.3.2 The Role of Stalk Binding Antibodies in Seasonal Infection

The prevalence and action of stalk binding antibodies in seasonal infections is not well understood (Section 4.4), and focus of study has so far been given to broad-binding examples, which appear to be extremely rare. A greater understanding of less broadly binding types – both their prevalence and their action – for example whether or not they are neutralising, and whether or not they lessen the virulence of infection – is needed in order to build a better understanding of the role of stalk binding antibodies and their effectiveness *in vivo*.

6.3.3 Determination of Physiological Concentrations of Antibody

The current, limited, understanding of physiological levels of antibody suggests that broad spectrum stalk-binding anti-HA antibodies identified to date may not be present in sufficiently high concentrations *in vivo* to have a neutralising effect (Section 5.5). Further studies are needed

in order to understand whether the binding strengths being observed would be sufficient to provide protection in humans.

6.3.4 Mutagenesis

The mutagenesis studies carried out in the studies of broadly-binding stalk antibodies, while extensive, do not cover all possibilities for escape (Section 5.5). It would be useful to attempt to breed escape mutants against antibodies FI6 and CR9114, and to attempt to culture escape mutants against antibodies D8/F10/A66 using additional base strains (Table 5.6). Such escape mutants, if bred successfully, would be helpful in understanding the critical residues of the epitope. Ideally one would wish, through mutation analysis or some other means, to understand the binding energy of each member of the epitope. This would enable a much more informed analysis of the likelihood of escape. If such an analysis proves impractical in the laboratory, computational methods may offer an alternative approach (Section 6.2.2).

6.4 The Changing Nature of Sequence Data

At the time of writing, the Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU) holds 9,358 sequences for HA for strains isolated between 1st January 2009 and 31st December 2009: that is, over twice the number of isolates it holds for H1N1 and H3N2 combined in all years prior to 2009. The numbers have since fallen substantially: a total of 1387 human HA sequences (across all subtypes) were deposited for 2010, and 899 for 2011. The peak was undoubtedly caused by increased surveillance following the emergence of the pandemic: nevertheless, overall, the number of sequences added annually is likely to trend upwards, following the same trend that has been observed since the early 1990s (Figure 5.1).

While the increasing volume of data poses a computational challenge, it will provide new opportunities. It remains to be seen how much the volume will, of itself, help in the overall understanding of the mechanisms which allow apparently simple behaviour, such as antigenic behaviour, to emerge from the overall complexity of a system consisting of a virus that interacts, at any one time, with tens of millions of hosts (Bedford, Rambaut, and Pascual, 2012). It may well exacerbate problems of phylogeny, which are touched on in Appendix A. Bush et al. (1999) identified an excess of mutations on the terminal branches of the phylogenetic tree, and recommended that studies of positive selection should be confined to internal branches. This, and other methods that focus on the mainstream evolution of the virus, may become increasingly necessary as the volume of data increases.

The increasing volume of data may help us to identify subtler patterns underneath the apparently simple progression of influenza's evolution today – for example a prevalence of particular strains in particular geographies, propensities linked to host genetic factors, and the presence of sub-populations of subtypes. It should enable the rather general descriptions of dominant strains worldwide (see Section 2.4.1) to be replaced by a more meaningful and numerically precise account. However, if such an account is to include descriptions of antigenic relatedness, a computational model of such relatedness will be needed, given the very large number of sequences that will be involved.

Deep sequencing offers the possibility of studying viral evolution within the host as well as between hosts, as we do now. Whether or not influenza follows the quasi-species model is contentious (Holmes, 2009), nevertheless it seems likely that the virus will adapt to the host environment once infection is established, and study of such adaptions should provide interesting insights into viral and host mechanisms. Variation of adaption between hosts could provide indications of genetic determinants of virulence. Intra-host transmission bottlenecks may help elucidate the determinants.

Appendix A - An Investigation of Nucleotide Mutation Rates in Influenza A H3N2 Haemagglutinin

A.1 Introduction

Our recent research has utilised a method developed by Shih et al. (2007) to infer selective pressure at a single amino acid site by identifying frequency switches: a method which I have extended to include subtler signs of pressure such as transient polymorphism, and applied to infer co-ordinated selective pressure amongst a number of amino acid sites in the HA2 polypeptide chain of Influenza A HA, as described in Chapter 5.

Researchers in many fields, including that of influenza evolution, have developed techniques to infer the action of selective pressure by comparing the ratio of non-synonymous to synonymous nucleotide substitutions (dN/dS). The large number of sequences available for influenza HA has made it a common target for such work. Typically, in such an approach, a nucleotide substitution model is calibrated against the history of the protein. A phylogenetic tree is inferred in order to account for intermediate ancestral states not represented in the available samples. The ratio of non-synonymous to synonymous nucleotide substitutions is then calculated, and compared to the value expected from the substitution model. A significantly high value implies positive selective pressure, while a significantly low value implies negative selective pressure.

Traditionally, for reasons related to calculation time and sample size, the calculation is aggregated over a number of sites and along all branches of the phylogenetic tree. Hence selective pressure at a small number of individual sites, or along a single branch of the phylogenetic tree, may be missed. Methods do exist that attempt to overcome these restrictions, and a useful account may be found in Chapter 8 of Yang (2006).

Existing Studies

Table A.1 summarises methods broadly based on this approach which have been applied to influenza H3 haemagglutinin and attempt to identify individual residues under selective pressure or to distinguish between pressure in antigenic and non-antigenic regions.

Study	Method	Extent
Ina and Gojobori (1994)	Analysis of antigenic site members vs. non-members	HA1
Fitch et al. (1997)	Single site analysis. Tree reconstructed by parsimony, changes counted at each site along each branch of the tree	HA1
Suzuki and Gojobori (1999)	Same approach and data set as Fitch et al. (1997) but a more stringent criterion for positive selection	HA1
Bush et al. (1999)	Single site analysis. Differentiation between mutations on terminal and nonterminal branches of the phylogenetic tree.	HA1
Yang (2000)	Tree and substitution model inferred by Maximum Likelihood (ML)	HA1
Suzuki (2004)	Grouping of amino acid locations via three dimensional window analysis utilising information on protein structure	HA1
Suzuki (2006)	Single site analysis	Entire genome
Wolf et al. (2006)	Analysis of antigenic site members compared to other regions	Entire genome
Suzuki (2008)	Analysis of antigenic sites A-E along grouped phylogenetic tree branches	HA1
Kosakovsky Pond et al. (2008)	Tree inferred by ML. Evidence is then explored for a directional substitution model which favours substitutions towards a specific amino acid at a particular site	Entire gene segment
Suzuki (2011)	Analysis of locations grouped around glycosylation sites	Entire gene segment
Chen and Sun (2011)	Parallel single-site analysis from multiple datasets	HA1
Murrell et al. (2012)	Single site analysis, episodic pressure	HA1
Tusche et al. (2012)	Detection of patches containing close residues with evidence of positive selection	Entire gene segment
Suzuki (2013)	Single site analysis, correcting for energetically unstable amino acid mutations	Entire gene segment

Table A.1: Published Analyses of H3 sequences using nucleotide substitution methods

In viral surface proteins, evolution at a single amino acid site is rapid and transient, as evidenced by the rapid changes observed in amino acid frequency plots. Single-site models may lack the sensitivity to identify changes of this nature. To address this, sliding window analyses have been developed that consider changes in locations closely located in the three-dimensional structure and hence can relax the criterion for positive selection at any single location. Suzuki (2004), employed windows of fixed size. Tusche et al. (2012) searched for continuous patches with a heightened dN/dS ratio, testing this approach on H3 HA1 and HA2 among other proteins.

A.2 HA2 Results from Published Studies

Approaches based on dN/dS are known to be conservative (Suzuki, 2013). Typically a dN/dS analysis will identify a small set of locations under positive or negative selective pressure, and a large number for which the analysis is inconclusive.

Study	H3 HA2 sites/regions under positive pressure
Suzuki (2006)	None
Kosakovsky Pond et al. (2008)	1
Suzuki (2011)	None
Tusche et al. (2012)	None
Suzuki (2013)	None

Table A.2: HA2 sites reported in those analyses from Table A.1 that cover HA2. None of these analyses list sites determined to be under negative selective pressure, although Figure 1 in Suzuki (2006) suggests that approximately 10% of HA2 locations satisfy the 95% likelihood criterion for negative selection.

Of the studies considered above, five cover H3 HA2 (Table A.2). Kosakovsky Pond et al. (2008) identified one HA2 location under pressure, with a whole-segment numbering of 347 corresponding to an HA2 numbering of 2. Tusche et al. (2012) do not identify any HA2 patches. While the latter result is perhaps unsurprising given the relative sparsity of substitutions in HA2, it is interesting that the former study only identifies one of the fixations in HA2 as being in a location undergoing positive selective pressure (Figure A.1). It is possible that the analysis in that study is influenced by the reversals at locations 32, 46, and 121. The three single-site dN/dS studies covered HA1 only: furthermore, all three are based on the dataset of Bush et al. (1999) and therefore cover a period when comparatively limited samples were available.

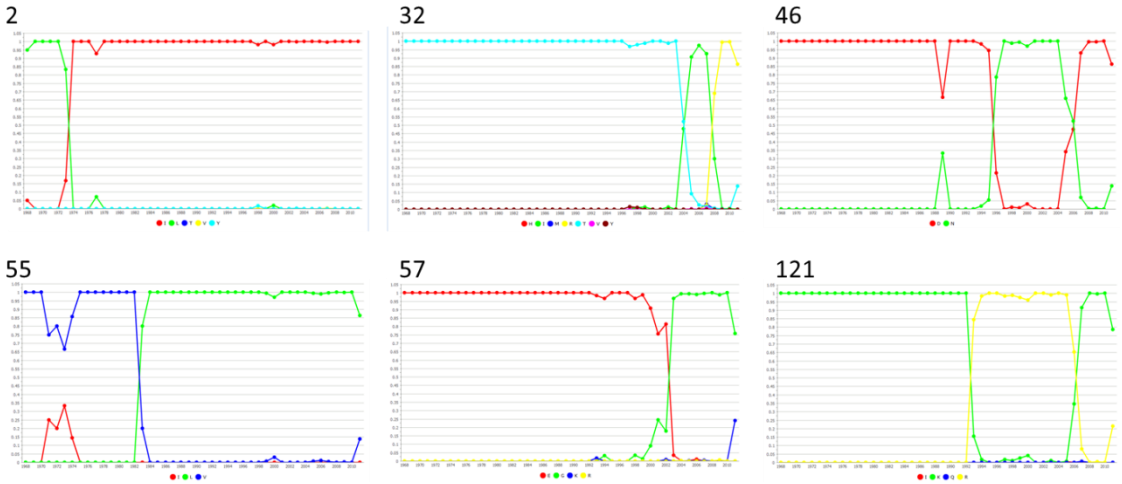


Figure A.1: HA2 locations undergoing fixations, of which only location 2 is identified as a location under positive selective pressure in Kosakovsky Pond et al. (2008).

A.3 An Investigation of H3 HA2 Selective Pressure Inferred From Nucleotide Substitution Rates

I decided to investigate positive selection in H3 HA2 using the single-site dN/dS calculation. I chose to run the calculation in strain samples isolated between 1999 and 2008, split into two groups (Table A.3). The period under study was chosen to match a period in which I had observed significant fixation activity (see Figure 5.10), and was split into two sets partly for reasons of computational economy and partly to reflect the division of activity observed in that period. It was anticipated that a focus on a relatively limited period in which specific directional activity could be observed in the fixation analysis would provide a clearer signal for the dN/dS calculations.

As an additional investigation, I repeated the directional substitution analysis of Kosakovsky Pond et al. (2008) focussing on the same period as above. Again it was anticipated that this focus might provide a clearer signal than the original broader-ranging analysis.

A.4 Methods

All available full-length human H3N2 HA nucleotide sequences were downloaded from the Influenza Virus Resource (Bao et al., 2008), and aligned with Muscle (Edgar, 2004). Sequences with gaps were eliminated. Sequences for HA2 were extracted, and duplicates removed (Table A.3). Selection analyses were performed in HyPhy v2.1.1 (Kosakovsky Pond and Frost, 2005).

Isolation Dates	**Number of unique HA2 nucleotide sequences**	**Number of unique HA2 AA sequences**	**HA2 Events inferred from frequency analysis**
June 1999 – May 2004	348	108	18, E57G, 123
June 2004 – May 2008	316	100	I32R, N46D, R121K

Table A.3: Number of sequences used in the HA2 studies, and the HA2 fixations inferred from frequency analysis in the periods under study. In the right-hand column, fixation events are shown as a transition between two amino acids (e.g. E57G) while polymorphisms are indicated purely by the location (e.g. 18)

For the dN/dS study, phylogenetic trees were estimated in HyPhy with the neighbour joining method, using the REV substitution model (Yang, 1994) and the TN93 distance model (Tamura and Nei, 1993) using global parameters. Nucleotide substitution biases were then estimated using the inferred tree by fitting the Generalised Time-Reversible model (Tavaré, 1986) with global parameters. dN/dS values were then estimated by HyPhy's Single Most-Likely Ancestor Counting (SLAC) method, applied to the whole tree at once, with ambiguities in reconstructed

codons averaged over possible codon states. This pipeline follows the approach recommended by the authors of HyPhy for a SLAC analysis.

HA1 sites were excluded from the above analysis in the expectation that this would allow a clearer signal to be obtained from the slower changing HA2 chain. To provide a comparison, a SLAC analysis was also conducted against full-length HA1 and HA2 sequences, using the above methods. The main reason for conducting this comparative analysis was a concern that that it might not be possible to obtain a representative phylogenetic tree from the HA2 sequences alone. The number of distinct full-length nucleotide sequences used in this analysis was 787 for the period 1999 –2004 and 593 for the period 2004 –2008.

For the directional substitution analyses, the nucleotide sequence of A/Hong Kong/1/1968 was added to each collection. Amino acid sequences were inferred from the nucleotide sequences and duplicates removed. Phylogenetic trees were then estimated from the remaining corresponding nucleotide sequences using the neighbour joining method as above, and the trees were rooted on A/Hong Kong/1/1968.

A.5 Results

SLAC Analysis of HA2 Sequences

At the default confidence level of $p < 0.05$, the SLAC method did not identify any sites under positive selection in HA2. 43 sites in the period 1999-2004 and 39 in the period 2004-2008 were identified as under negative selection: 17 of these were identified in both periods.

Results for sites of interest from the fixation analysis in Chapter 5 are shown in Table A.4

.

Site	1999 to 2004						2004 to 2008						1999-04 sel	2004-08 sel	Events
	Observed S Changes	Observed NS Changes	dS	dN	P{S leq. observed}	P{S geq. observed}	Observed S Changes	Observed NS Changes	dS	dN	P{S leq. observed}	P{S geq. observed}			
18	1	5	1.29	2.25	0.51	0.83	2	5	2.03	2.48	0.58	0.73	neut	neut	Polymorphism between 2000 and 2003
32	1	2	1.00	1.00	0.74	0.70	2	12	5.40	4.57	0.75	0.53	neut	neut	Fixations in 2004 and 2009
46	4	0	5.52	0.00	1.00	0.00	3	9	5.54	3.66	0.84	0.37	-ve	neut	Fixation in 2006
57	2	5	2.37	2.50	0.65	0.66	2	2	2.03	1.00	0.89	0.40	neut	neut	Fixation in 2002/2003
121	3	6	3.31	2.99	0.70	0.57	3	5	3.26	2.66	0.75	0.52	neut	neut	Fixation in 2006/2007
123	2	3	2.90	1.34	0.91	0.34	8	4	11.72	1.77	1.00	0.00	neut	-ve	Polymorphism in 2000-2002

Table A.4: Results of single-site SLAC analysis of HA2 sequences isolated in the two periods studied, for events of interest identified from fixation and polymorphism analysis. For each of the two periods, the following information is provided:

Observed S/NS Changes – a count of silent and nonsilent changes observed at the site

dS, (dN) – The number of observed synonymous (non-synonymous) substitutions per synonymous (non-synonymous) site, as calculated by the SLAC model

P{S leq observed) – The probability (calculated from a continuous extension of a binomial distribution) of observing this or a greater number of synonymous changes on the assumption that the selection is neutral – this is the *p*-value of positive selection at the site

P{S geq observed) – As above, the *p*-value of negative selection at the site

1999-04, 2004-08 selection – Selection applying at the site (neutral or negative)

Fixation or Polymorphism Events – Events as inferred from amino acid frequency analysis

The results in Table A.4 suggest a lack of predictive power compared to the amino acid frequency analysis. The two models are not in conflict in the sense of one predicting negative selection where the other predicts positive selection, but the SLAC analysis appears to have less sensitivity to the presence of positive selection.

Period	Location	Observed S changes	Observed NS changes	P{S leq. observed}
1999-2004	18	1	5	0.51
	27	3	5	0.69
	32	1	2	0.74
	50	1	2	0.74
	57	2	5	0.65
	94	0	1	0.73
	97	0	1	0.75
	98	1	4	0.46
	100	0	1	0.67
	121	3	6	0.70
	143	0	1	0.72
	147	0	6	0.09
	150	2	5	0.58
	160	0	3	0.64
	163	0	1	0.69
	172	0	3	0.35
2004-2008	18	2	5	0.58
	32	2	12	0.75
	71	1	2	0.69
	84	1	2	0.74
	99	0	1	0.67
	101	1	2	0.74
	104	0	1	0.77
	121	3	5	0.75
	139	0	1	0.72
	155	1	2	0.68
	168	0	1	0.77
	172	0	4	0.25
	173	0	2	0.44

Table A.5: Locations in each examined period that were assigned a p-value for positive selection less than 0.8. Locations at which the amino acid frequency model identified events in the period are shaded.

Table A.5 lists all locations with a p-value for positive selection < 0.8: this includes 4 out of the 6 locations identified by amino acid frequency analysis, along with a further 25 hits. Many of

these have low observed N and S changes (Figure A.2). Interestingly, the four locations identified by amino acid frequency analysis have relatively high observed N and S changes, the lowest total being 6.

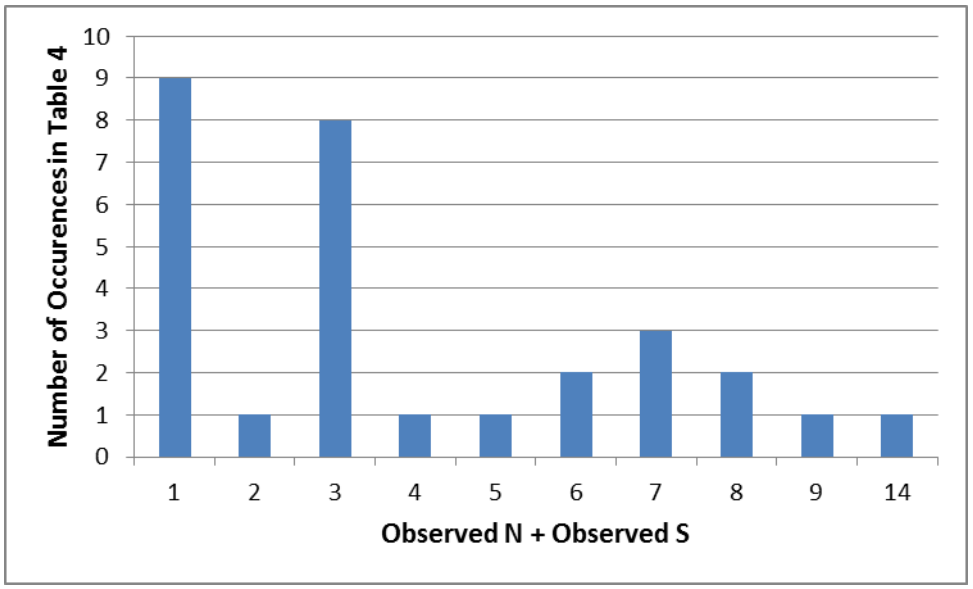


Figure A.2: Frequency of the total observed N and S counts seen in the results listed in Table A.5.

Of the results listed in Table A.5 which are not identified in the amino acid frequency analysis, the two with the highest total observed N and S counts are 121 in 1999-2004 (9) and 27 in 1999-2004 (8). While location 121 underwent fixations in 1992/1993 and 2006/2007 it displays only a low degree of polymorphism in 1999-2004. Location 27 shows little polymorphism in the period. Likewise the two locations with the lowest p-values for positive selection: 147 in 1999-2004 (0.9) and 172 in 2004-2008 (0.25) display only modest polymorphism (Figure A.3).

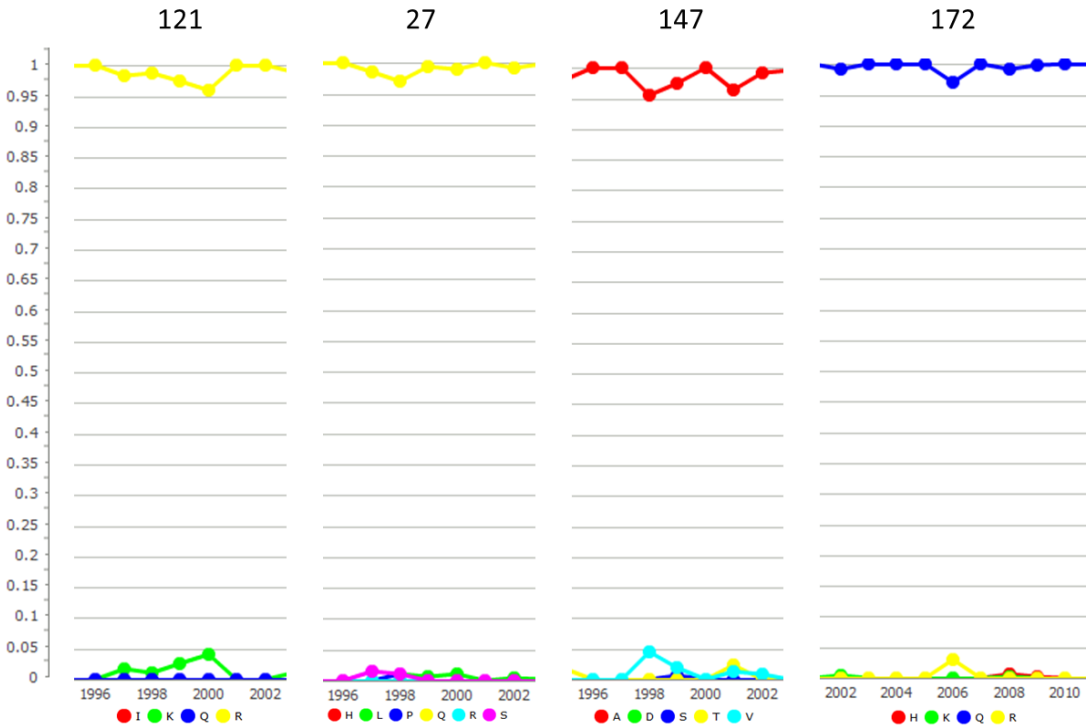


Figure A.3: Amino acid frequency charts for four residues identified by SLAC analysis as being under positive selection in the periods shown.

SLAC Analysis of Full-length HA Sequences

Once again, at the default confidence level of $p < 0.05$, the SLAC method did not identify any sites under positive selection in HA2. 6 locations in HA1 were identified as being under positive selective pressure in the period 1999-2004 and 9 in the period 2004-2008. (Table A.6).

Period	Location	P{S leq. observed}	Antigenic Site
1999-2004	45	0.01	C
	50	0.01	C
	92	0.04	E
	220	0.0	-
	226	0.05	D
	229	0.00	D
2004-2008	3	0.01	-
	50	0.04	C
	53	0.00	C
	142	0.02	A
	157	0.01	B
	173	0.03	D
	194	0.01	B
	223	0.03	-
	261	0.02	E

Table A.6: HA1 locations identified by SLAC analysis as being under positive selective pressure

For HA2 sites of interest identified from frequency analysis, counts of silent and nonsilent changes obtained from the full-length SLAC analysis are broadly in line with those obtained from the HA2-only analysis (Table A.7), indicating that the phylogenetic trees are similar. From the point of view of inferring sites under selective pressure, neither method seems preferable: however it would be worthwhile to conduct a more detailed examination of the inferred phylogenetic trees for HA2 and full-length HA, for example using bootstrap analysis to examine stability.

A.6 Directional Evolution Study of Full-Length Sequences

The 2008 study published by Kosakovsky Pond et al. was based on full-genome sequences of strains isolated between 1968 and 2005. For HA, 284 unique full-length sequences were available. Software used for the study was provided publicly in version 1.0 of the HyPhy software package (www.hyphy.org). The sequence datasets used in the study are published.

The analysis requires a rooted tree, and for this purpose the sequence of A/Hong Kong/1/1968 was included in the 2000 and 2004 sequence sets. Analysis of the HA2 sequence sets using the provided H5N1 substitution model in the current version of HyPhy, 2.1.1, yielded no results.

Site	1999 to 2004						2004 to 2008					
	HA2-only Analysis			Full-length Analysis			HA2-only Analysis			Full-length Analysis		
	Observed S Changes	Observed NS Changes	P{S leq. observed}	Observed S Changes	Observed NS Changes	P{S leq. observed}	Observed S Changes	Observed NS Changes	P{S leq. observed}	Observed S Changes	Observed NS Changes	P{S leq. observed}
18	1	5	0.51	1	6	0.44	2	5	0.58	2	6	0.47
32	1	2	0.74	1	2	0.74	2	12	0.75	1.5	14.5	0.51
46	4	0	1.00	4	0	1.00	3	9	0.84	3	4	0.92
57	2	5	0.65	2	6	0.49	2	2	0.89	2	2	0.89
121	3	6	0.70	2	6	0.52	3	5	0.75	1.9	5.5	0.58
123	2	3	0.91	2	1	0.97	8	4	1.00	8	4	1.00

Table A.7: Comparison of Silent and nonsilent changes, and derived p-value of positive selective pressure, for identified sites of interest, showing values obtained both from the HA2-only SLAC analysis and the Full-length SLAC analysis.

Site (Segment numbering)	Chain	Site (H3 Numbering)	Antigenic Site	Preferred AA	Estimated Bayes Factor		
					Published Figures	My Figures	
						H5N1 model	REV model
5				I		2695	639
10				I	>100000	704848	23875
32	HA1	16		G		1092	693
48		32		D		2427	
61		45	C	S	338		
74		58		I		360	385
94		78	E	G		3525	343
101		85		D		945	
151		135	A	T	2109		
161		145	A	K	1936		
171		155	B	H	3047		
174		158	B	K	212		
190		174	D	F		4040	
191		175	D	D		415	
230		214	D	I		320	594
233		217	D	I			135
236		220		R		>100000	
245		229	D	R	>100000	>100000	
246		230	D	I		714	438
252		236		I		179	196
264		248	D	T	1112		
347	HA2	2		I	102		
395		50		G		873	264
434		89		I		179	195
453		108		I		1683	1220
477		132		D		417	
483		138		F		3178	
490		145		D		239	
494		149		I		711	435
505		160		D		238	
520		175		G		187	296
566				I		118	130

Table A.8: Sites identified as undergoing directional evolution identified using the method of Kosakovsky Pond et al. (2008). The Estimated Bayes Factor is returned by the model, and is a measure of the power of a directional model favouring the preferred amino acid compared to a neutral model.

I attempted to reproduce the results originally published by Kosakovsky Pond et al. for H3N2 HA in order to confirm that I was running the analysis correctly, but obtained quite different results (Table A.8). In correspondence with the author, it transpired that the results were obtained with the REV substitution model rather than H5N1. Using this model, I obtained a close subset of my earlier results. It is striking that, across these two results sets, 11 of the 25 identified sites have leucine as the preferred residue.

A.7 Discussion

The inference of selective pressure from synonymous and nonsynonymous nucleotide substitution rates has, since its inception (Miyata and Yasunaga, 1980), become sufficiently ubiquitous that it is tempting to regard a divergence in substitution rates as the definitive indicator of evolutionary pressure. As evidenced in this study, however, the application of the technique to the analysis of single site mutations in a rapidly changing organism such as influenza presents several difficulties.

Firstly, even with the high degree of change seen in influenza haemagglutinin, and with the large number of samples available, the number of changes at any one site is small, and the variation in the number of silent changes comparatively large. For example in the five year HA2 samples used in this study, in the sites identified in Table A.4, the number of silent changes at a site ranges from 1 to 8, while the number of nonsilent changes ranges from 1 to 12, with 12 being a notable outlier. In such circumstances, it is difficult to separate signal from noise and the predictive power of the approach becomes limited.

While it may be possible to improve the signal-to-noise ratio by sampling over a larger time period, such an approach will exacerbate another problem encountered in this study, which is that selective pressure at a site may change over time. Consider the 'directional sweep' exemplified by site 2 in Figure A.1, in which isoleucine is completely replaced in the population by leucine. As noted by Kosakovsky Pond et al. (2008), selective pressure at such a site may be highly negative in the period before the change (when isoleucine dominates) and in the period after the change (when leucine dominates) but is highly positive during the change itself. Averaging over too long a time period is likely to miss events such as this one, which occurs over a short time period in a manner characteristic of changes in influenza HA.

The directional evolution technique of Kosakovsky Pond et al. was introduced to detect just such switches, by focussing on directed change at the amino acid level. This method, however, has drawabacks of its own. Because it works at the amino acid level, the small numbers problem is even more acute. Also, if the site experiences reversals during the sample period (such as those seen in sites 46 and 121 in Figure A.1), both switches are likely to be missed due to averaging. Kosakovsky and Pond identify 7 sites in HA1 undergoing directional evolution compared at the 63 sites identified by Shih et al. (2007) in the same period. While it is possible that not all the switches were apparent in the 284 samples used by Kosakovsky Pond et al.

compared to the 2,248 samples used by Shih et al., reproducing the directional evolution analysis on a four node high performance compute cluster took approximately five days of computational time whereas running the Shih analysis on my database of over 10,000 sequences completes in a few seconds and is run on-the-fly in the course of requesting a page on the web site.

The techniques applied in Chapters 3-5 aggregate effects across all branches of the phylogenetic tree. This highlights an important consideration in the study of the phylogeny of highly mutable organisms such as viruses as opposed to the study of larger and more stable organisms such as plants or animals. The samples we collect of the former, particularly under the intense scrutiny of surveillance programmes such as that in place for influenza, map out evolutionary 'dead ends' throughout the history of the phylogenetic tree. Over the time period for which data exists, we have not witnessed the evolution of differentiated organisms via genetic drift. These factors account for the topology of the tree, with wide branches and a single mainline trunk. In contrast, when we examine the phylogeny of more stable organisms, we typically only have samples of the evolutionarily successful 'endpoints', of which there may be many. We can infer common ancestors, but we will typically have no sight of evolutionary dead ends that occurred in the past. These factors influence the questions of interest and the approach that should be taken. For example, in the case of a virus, if we are interested in directed evolution of the organism over a representative period, we may wish to study evolution along the main branch alone, as in Chapter 3.

Another issue observed with viral samples is the presence of inter-host adaptive mutations, for example the substitutions that are likely to occur if mammalian influenza is grown in eggs, or mutations observed in samples from severely ill patients where the virus has adapted to culture in lung tissue. Such mutations are not linked to a specific evolutionary pattern, and may be found in samples from many branches of the phylogenetic tree. A current example is in sites 153 to 157 of H1N1pdm HA1, which have been seen to vary in some samples since the beginning of the pandemic. The cause of this variation is unknown, but it occurs in phylogenetically distinct samples in many branches (McCauley et al., 2012). Such adaptive mutations will be given prominence by mutation rate analyses, due to their presence in multiple branches, and this may account for the identification of sites such as those highlighted in Figure A.3.

Suzuki (2013) identifies a further problem with the dN/dS approach, which is that, in calculating the ratio, all amino acid locations reachable by a single nucleotide mutation are considered physiologically plausible. In practice, some will not be, and the discrepancy means that the approach will under-estimate the significance of such substitutions as are observed. He attempts to address this by removing from consideration substitutions that would violate a free energy criterion. While this does allow a greater number of positively selected locations to be identified, there are problems with the approach. Firstly, the free energy calculations and the free energy threshold for each location have to be derived via approximations from the substitution history available for the location, and from the limited number of crystal structures available for the protein. Secondly, the approach will not take account of functional limitations on allowed substitutions, which, in the case of functional regions of a protein, such as the RBS and fusogenic region of HA, may impose additional limitations that will be missed by consideration of free energy calculations alone.

dN/dS predictions of selective pressure are based on the assumption that synonymous codon mutations are not subject to selective pressure, however codon bias has been identified in influenza A PB2 (Marsh et al., 2008) and in other RNA viruses (Holmes, 2009). The argument has been put forward that selective pressures relating to plasticity are at work in HA, in addition to widely acknowledged host pressure relating to tRNA abundance and other host factors (Plotkin and Dushoff, 2003). The fundamental assumption of synonymous mutation neutrality is therefore questionable.

Initial results from directional evolution analysis suggest that this method loses power when confined to a small run of years, possibly because such a sample does not provide enough data to accurately calibrate the substitution model. The SLAC analysis, on the other hand, appears to perform well with a five year window. One possible avenue to explore is a time-based sliding window SLAC analysis, which could reveal variation in selective pressure at individual sites when events such as frequency switches occur. A concern with this approach is the stability of a phylogenetic tree derived from samples from a small number of years. Bootstrapping techniques could be utilised to investigate the stability of such trees and hence derive an understanding of the minimum window period that is likely to produce acceptable results.

Bootstrapping of phylogenetic trees would also provide an understanding of the most stable branches. Confining analyses (SLAC, directional selection and frequency analyses) to these branches may eliminate noise sourced from adaptive mutations and evolutionary dead ends.

Appendix B – Software Used In This Study

B.1 Overview

In previous work (Lees, 2009), I created a database of H3N2 sequences and HI assay information, collected from published sources. I created a web-based front-end through which consolidated information could be viewed: for example summarizing all assays relating to a particular strain.

For this work, I extended the database to cover H1 strains as well as neuraminidase N2 sequences, and different assay types, such as neutralisation, and HI in the presence of oseltamivir. In addition, I developed a 'visualisation workbench', which can be used to explore the spatial distribution of mutations, fixation events and mutation clusters. I include in this chapter an analysis of sequence and assay data quality, and a brief explanation of the functions and rationale of the web site.

B.1.1 Scientific Aims

The database and web site were developed to address the following scientific objectives:

> *To provide a representative set of H1, H3 and N2 sequences covering the recorded period of activity*. While large, publicly available sources exist, none are comprehensive, particularly in coverage of strains isolated before 1990. For this work, I wished to assemble as representative as possible a set of sequences covering the recorded active period of those subtypes.
>
> *To improve sequence quality and matches between sequences and assay information through consensus and curation.* In the course of assembling the sequence data, I encountered classification issues, naming issues (for example inconsistent abbreviation) and a small number of low-quality sequences. By developing my own database, I could bring together sequences from various sources, and pay careful attention to classification and consensus.
>
> *To provide a consistent, curated set of HI assay information*. In contrast to the sequence data, no publicly accessible database of HI assay data exists. By consolidating assays from numerous published sources, I aimed to create a significantly larger body of data than has previously been used for modelling purposes.

To facilitate the study of HA sequence evolution. While many general-purpose sequence analysis tools exist, I wished to make it easy to identify differences between specific sequences, or trends across the entire population, by referring conveniently and directly to the underlying sequence database.

To visualise evolutionary trends on the structure. I wished to examine the location of evolutionary events such as substitutions and fixations on the protein structure, in a way that would facilitate interactive discovery and hence build knowledge of patterns and trends.

Pages in addition to those provided on the website today were developed during the course of the research, for example to test particular techniques for substitution cluster identification, to classify substitutions in particular ways, and to test computational models linking antigenic and sequence development. The website as it stands includes those pages that I feel are of general interest, or which support specific results described in this work.

B.2 The Web Site and Database

Here I briefly describe the functionality offered from the home page of the web site (Figure B.1). The pages listed under ‘About the Database’ and ‘Basic Reports’ were first developed for my previous work, while the remaining pages were developed for this work.

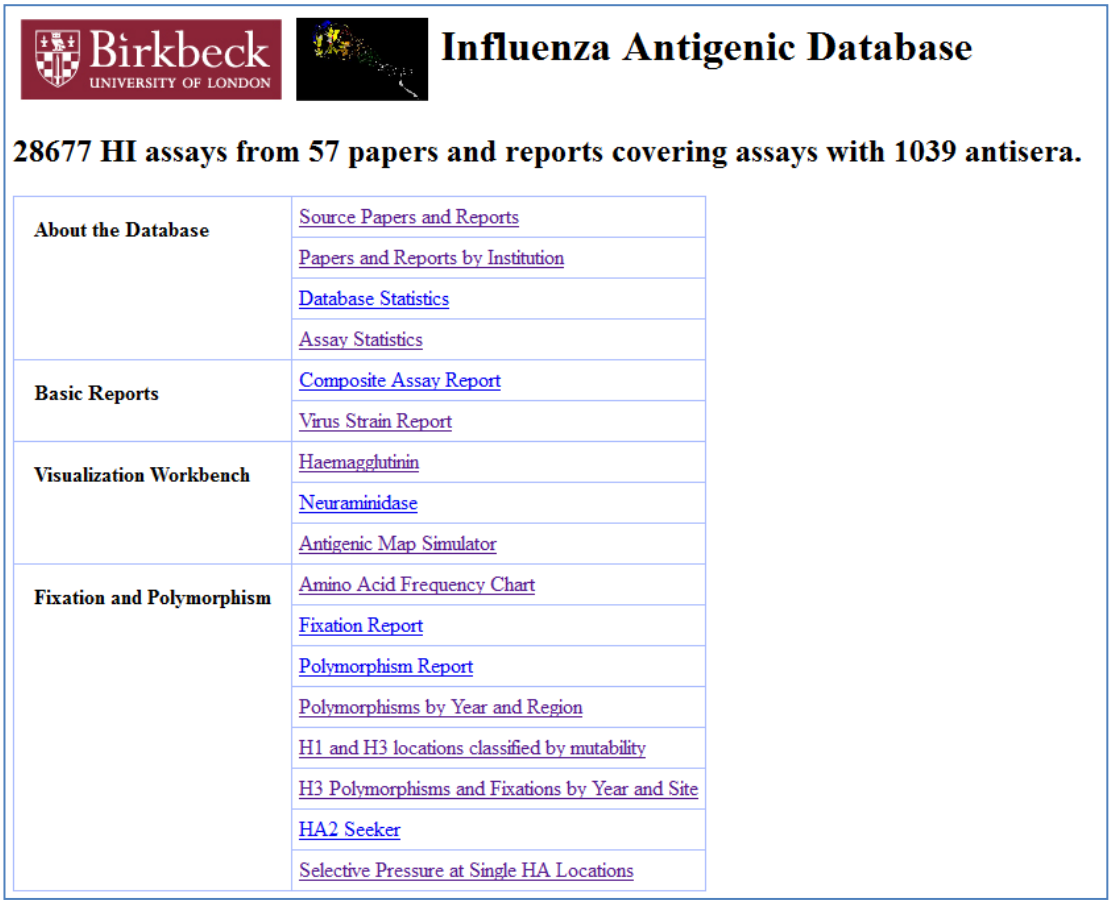


Figure B.1: The Web Site Home Page

B.2.1 About The Database

The *Source Papers and Reports* and *Papers and Reports by Institution* pages provide details of all sources from which HI data has been extracted. The individual reports are clickable, providing a summary of the data extracted from each one.

The *Database Statistics* page provides an overview of the data held within the system. At the time of writing, this encompasses HA sequences for 15,826 unique strains, 9678 H3N2 and 6148 H1N1; and 28,677 HI assays, extracted from 57 papers and reports.

The *Assay Statistics* page (newly developed for this work) provides some statistics on commonly conducted assays, for which multiple results are held in the database (Figure B.2). These statistics are discussed in Section 2.1.

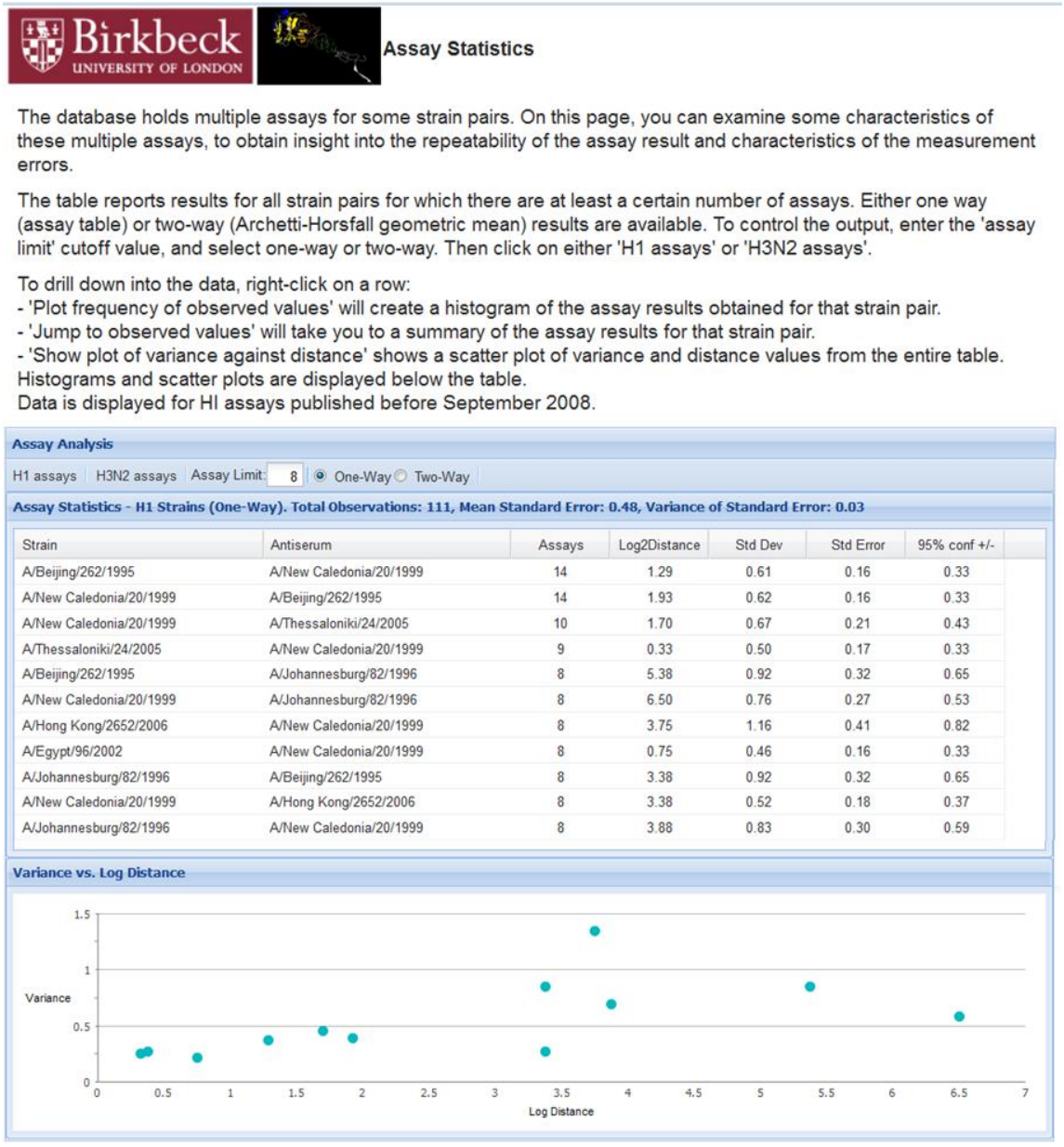


Figure B.2: Assay Statistics for frequently occurring strain pairs

B.2.2 Basic Reports

The *Composite Assay Report* merges HI assay data from all available reports for selected strains and antisera.

The *Virus Strain Report* provides information held in the database for the specified strain, including a summary if its inclusion in HI assays, consolidated assay results, and available sequences.

B.2.3 Visualisation Workbench

The web site embodies a flexible tool for visualising sequence differences on the HA molecule (Figure B.3). A similar facility exists for visualising N2 sequence differences. In the Compare Strains tab (shown in the figure), the user enters two strains of the same subtype. Type-ahead is provided, so that after typing the first few characters, matching strains with sequences present in the database are provided as a dropdown. Amino acid differences between the two strains are highlighted on a visualisation of the HA molecule rendered in Jmol (Hanson, 2010), using PDB structure 1RU7 (A/Puerto Rico/1934) (Gamblin et al., 2004) for H1 strains and 1HGD (A/Aichi/2/1968) (Sauter et al., 1992) for H3 strains. The structures chosen have good resolution (2.6 and 2.7 Å respectively) and were taken of the protein in complex with the sialic acid receptor. These were chosen in preference to structures of the protein complexed with antibodies, as the structural shape of the HA molecule in the latter structures may be more grossly affected by binding to the antibody, given its much larger size and surface of interaction compared to the sialic acid complex.

Through checkboxes, the user can elect to display a single monomer or the whole trimeric structure, can show or hide the HA2 chain, and can colour those locations comprising the RBS in yellow. Residues differing between the two strains are coloured red on the structure. Those which constitute fixations (i.e. the substitution observed in the later of the two strains becomes fixed across the whole population) can optionally be coloured in blue: here I take Shih's definition of an "effective switch" (Shih et al., 2007), which includes a statistical test of significance, as constituting a fixation. In addition, the entire facilities of Jmol (for example to select a different display style or to measure distances between residues) are available by clicking or right-clicking on the display.

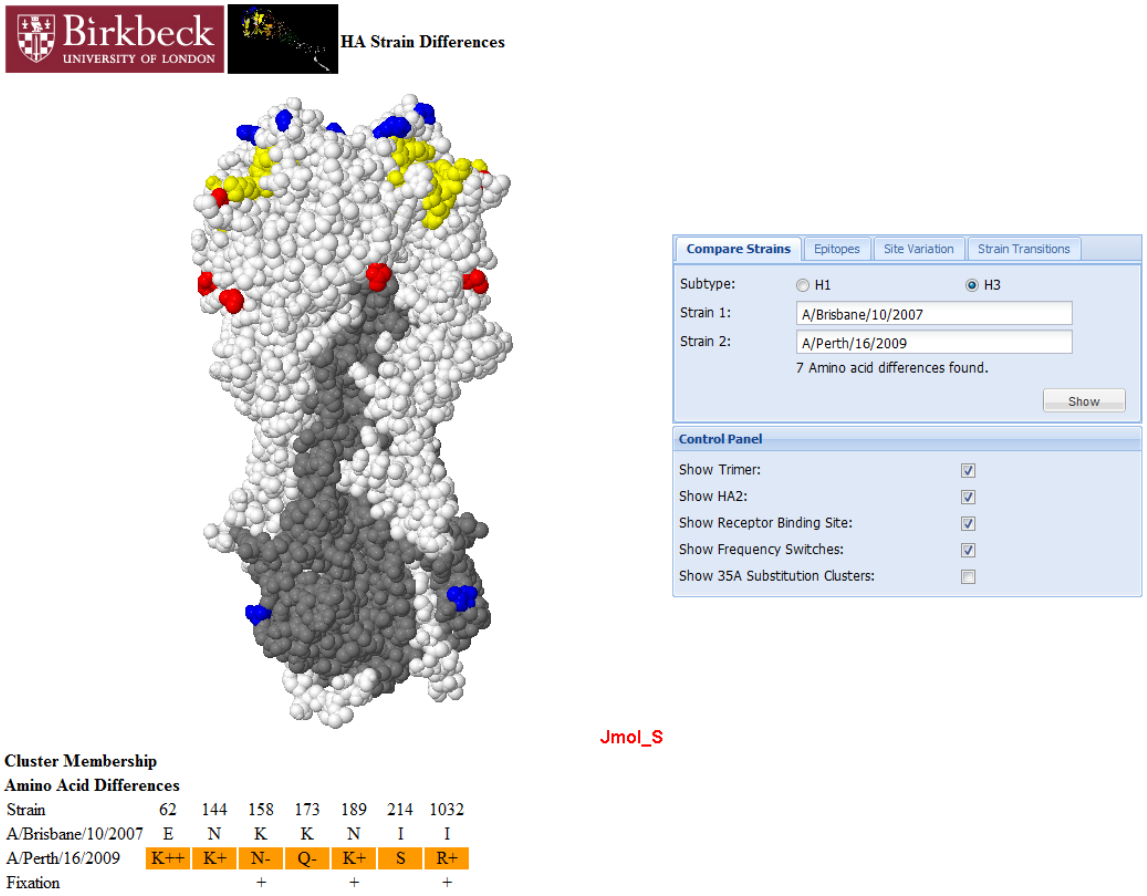


Figure B.3: Structural and Amino Acid comparison of HA strains from the web site (the Jalview amino acid sequence display has been omitted for brevity)

A further checkbox enables the display of substitution clusters. A substitution cluster is defined as a set of three or more substitutions where all members of the set are within a specified distance of each other. In this visualisation tool, the distance is set to 35Å. The biological significance of substitution clusters is discussed in Chapter 3.

Underneath the Jmol display is a quick summary of residue differences, including charge differences. Fixations are indicated. Below this (not shown in the figure) is a rendition of the two sequences in Jalview (Waterhouse et al., 2009) in which sequence differences are highlighted.

Other tabs on the screen provide additional visualisations superimposed onto the HA structure as follows (Figure B.4):

Epitopes – Indicates (in red) the B-cell epitopes from selected influenza A / antibody complexes in the PDB (the residues comprising the epitope are those identified in the published article accompanying the structure).

Site Variation – Shows visually the degree of polymorphism at each location in strains from the selected year, where white (for HA1) or grey (for HA2) indicate no variation, and shades from light yellow through to red indicate progressively greater degrees of variation. Fixations occurring that year are indicated in blue, if the 'Show Frequency Switches' option in the Control Panel is selected.

Strain Transitions – Shows substitutions and fixations occurring between successive dominant H3N2 strains. A particular strain transition can be selected, or the sequence can be cycled through with the arrow buttons.

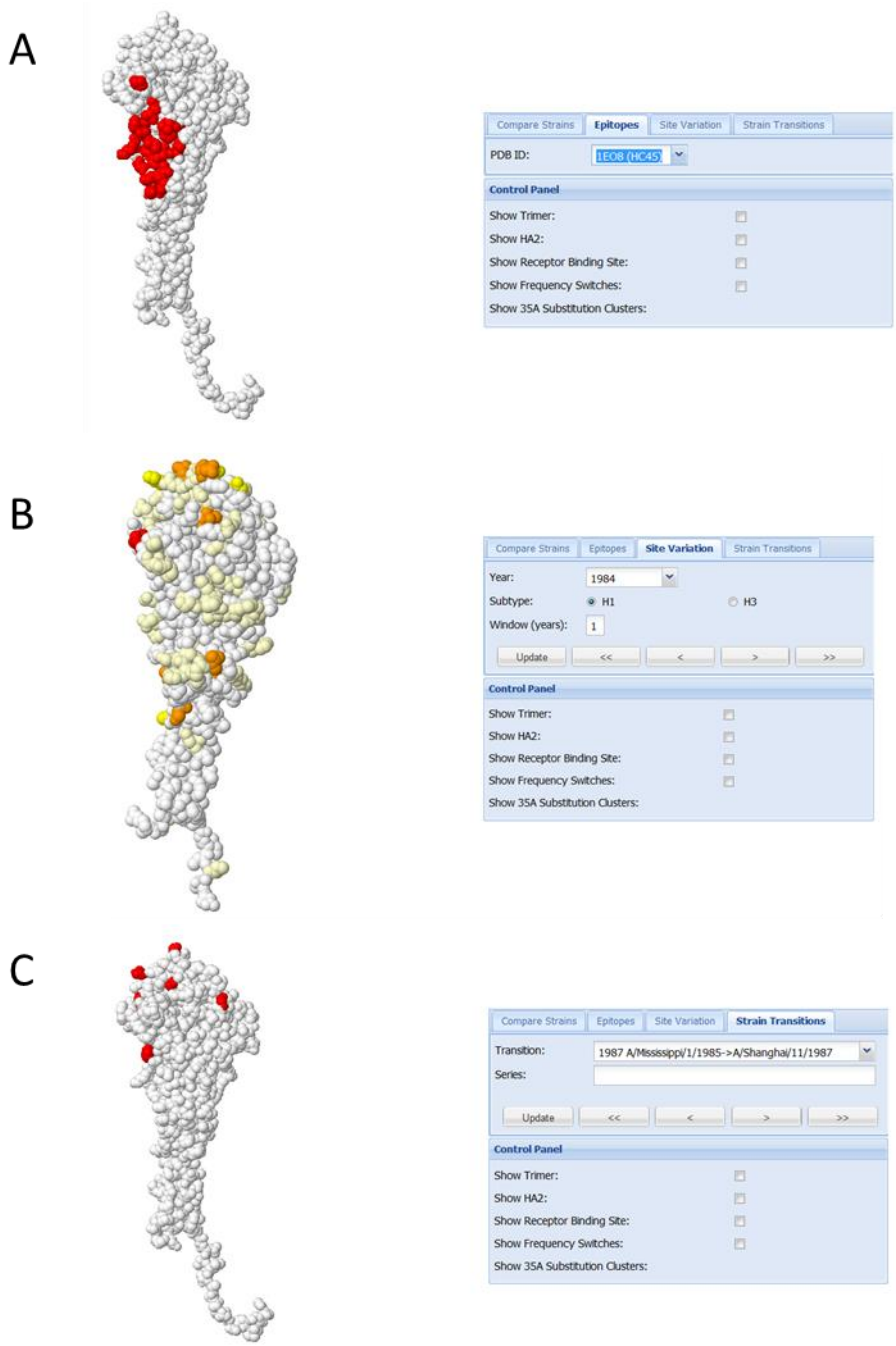


Figure B.4: Additional visualisations of HA available from the Visualisation Workbench**.** (A) – epitopes of selected structures; (B) – site variation heat map for a selected year; (C) – substitutions between two successive H3 dominant strains.

The *Antigenic Map Simulator* generates simulated antigenic maps based on mutation characteristics provided by the user. It is discussed in detail in Chapter 4.

B.2.4 Fixation and Polymorphism

The *Amino Acid Frequency Chart* provides a chart showing the relative frequency at which specific amino acids are observed at the specified location in each year for which samples are available, and a second chart showing the number of sequences available in each year (Figure B.5).

The *Fixation Report* lists the year and location of all fixations identified in the H1 and H3 sequence collections, using the method previously described by Shih et al. (2007). The report is sortable by column header and can be exported to Excel.

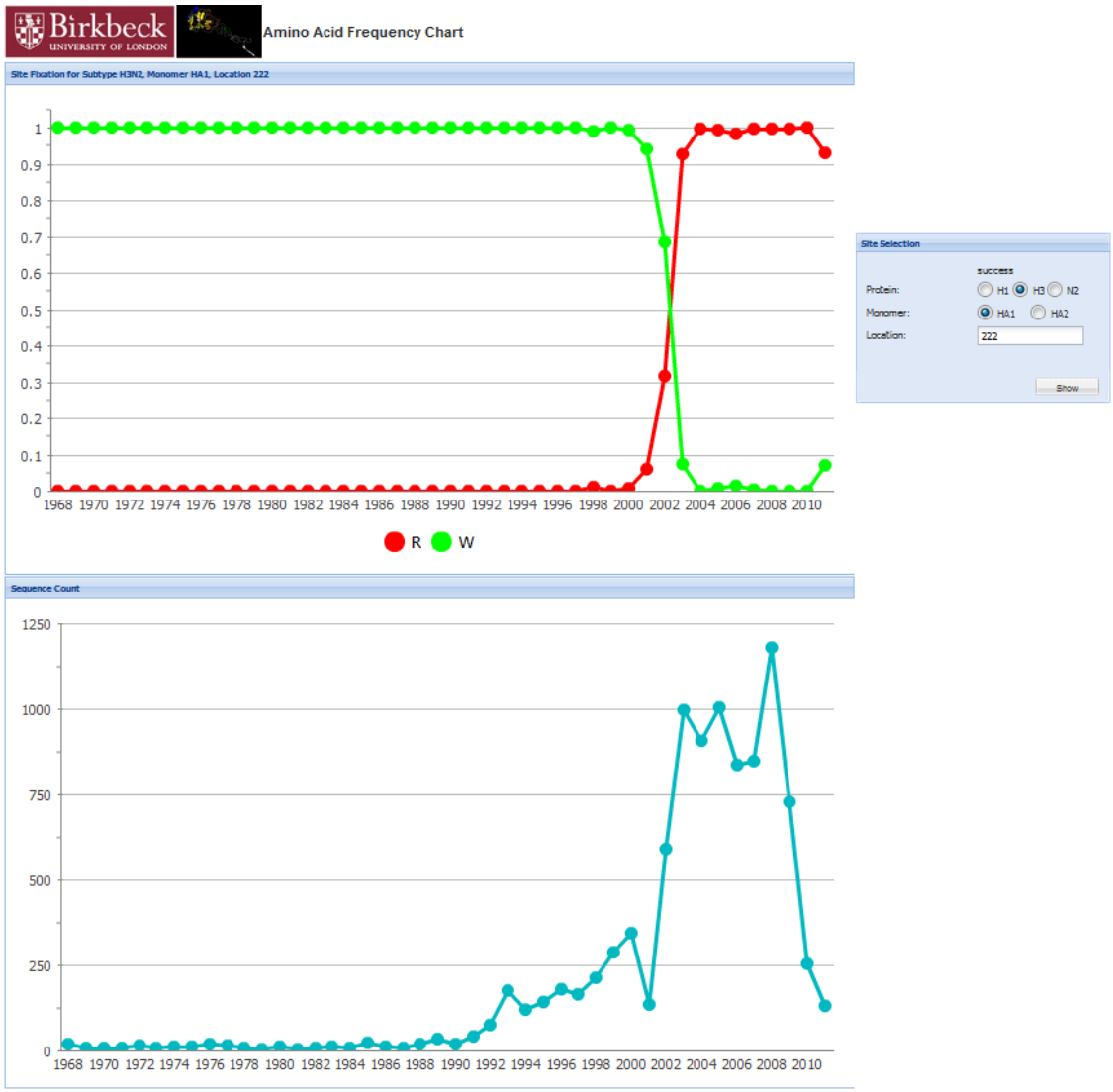


Figure B.5: Amino Acid Frequency Chart Web Page, showing the frequency chart for H3N2 HA1 location 222, and underneath a count of the number of H3N2 sequences available in each year.

The report indicates the perpendicular distance of the location at which the fixation occurs along the axis of the molecule (a precise definition is provided on the page). This distance can be used to identify fixations in the fusogenic region, using the definition of that region provided in Chapter 5.

The *Polymorphism Report* lists the location and year of polymorphisms in a format similar to the fixation report. Here a polymorphism is defined as a location in which, across all the sequences available for a particular year, no one amino acid type is present in more than a threshold percentage of sequences. The threshold can be set on the page, but defaults to 80%.

Polymorphisms By Year And Region generates a chart showing the number of fixations and polymorphisms occurring in each year in selected regions: antigenic sites A and B (as a single region), antigenic sites C and E (as a single region), and the fusogenic region (defined as in Chapter 5). The analysis is confined to H3 sequences. The plot can be restricted to particular data series of interest via the checkboxes (Figure B.6).

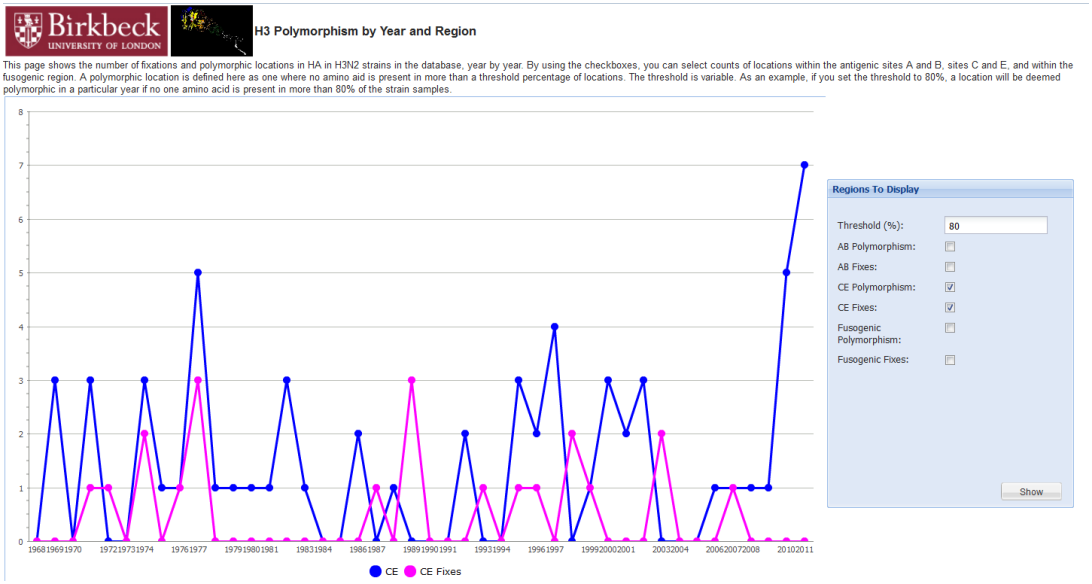


Figure B.6: Polymorphism chart for H3 sequences, showing in this case the number of polymorphisms and fixations in each year in antigenic sites C and E.

H3 Polymorphisms and Fixations by Year and Site tabulates the information displayed by *Polymorphisms By Year And Region* in a form convenient for import into Excel or other applications.

H1 and H3 Locations classified by Mutability classifies each location according to the mutability observed:

- Switchers are those locations which have undergone a fixation event;
- Polymorphic locations are those which have not, but have undergone a period of polymorphism as defined in the *Polymorphism Report*;
- Invariant locations are those which always have the same dominant amino acid, and where the dominant amino acid is present in at least 95% of samples each year;
- Intermediate locations are other locations not falling in to one of the above classifications.

The *HA2 Seeker* lists strains for which an entire HA2 sequence is available which are genetically similar to a nominated target strain. Its use is explained in Section 2.4.2.

Selective Pressure at Single HA Locations allows the user to compare an amino acid frequency analysis for a given H1 or H3 location with selective pressure estimates obtained from counts of nonsilent and silent mutations, using the Single Likelihood Ancestor Counting Method (Kosakovsky Pond and Frost, 2005). This analysis is discussed further in Appendix A.

B.3 Software Used in this Study

B.3.1 The Web Server

The web server is built on open source components and will run on Linux or Microsoft Windows. Software components of the web server are shown in Table B.1.

Name	**Purpose**	**Developed By**
Apache HTTP Server 2.2	Web Server	The Apache Foundation
PHP 5.4	Server-side scripting	The PHP Group
MySQL Server 5.1	Database	Oracle, Inc

Table B.1: Server Software Components

B.3.2 The Web Client

The web client is also built on open source components. It has been tested with Mozilla Firefox v20, Google Chrome v27, and Microsoft Internet Explorer v10. Adobe Flash and Oracle Java must be installed in the browser for full functionality. Software components are shown in Table B.2.

Name	Purpose	Developed By
Ext JS 3	Client-side scripting	Sencha, Inc
Jmol 13	Visualisation of protein structures	www.jmol.org
Jalview	Sequence visualisation	Waterhouse et al. (2009)

Table B.2: Web Client Software Components

B.3.3 Other Software

Other software used in the course of this study is listed in Table B.3.

Name	Purpose	Developed By
R 2.11	Statistical analysis	http://www.r-project.org/
Muscle	Sequence alignment	Edgar (2004)
HyPhy 2.1	Phylogenetic Analysis	Kosakovsky Pond et al. (2005)
Word 2010	Word processing	Microsoft, Inc
Excel 2010	Data manipulation	Microsoft, Inc
PSPad	Source code editing	Jan Fiala
Perl 5	Data manipulation	www.perl.org
Dropbox	Backup and replication	www.dropbox.com
Unfuddle	Source control and defect tracking	www.unfuddle.com

Table B.3: Web Client Software Components

Bibliography

Air GM. 2011. Influenza neuraminidase. *Influenza and other respiratory viruses* (November 16).

Air GM, Laver WG, Webster RG. 1990. Mechanism of antigenic variation in an individual epitope on influenza virus N9 neuraminidase. *Journal of Virology* 64, no. 12 (December): 5797–5803.

Akere O. 2012. Validation of antibody binding in the fusogenic region of certain H3N2 strains - Project Report for MSc Microbiology at Birkbeck, University of London: U.K.

Anders EM, Hartley CA, Jackson DC. 1990. Bovine and mouse serum beta inhibitors of influenza A viruses are mannose-binding lectins. *Proceedings of the National Academy of Sciences of the United States of America* 87, no. 12 (June): 4485–4489.

Archetti I, Horsfall FL. 1950. Persistent antigenic variation of influenza A viruses after incomplete neutralization in ovo with heterologous immune serum. *The Journal of Experimental Medicine* 92, no. 5 (November 1): 441–462.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. 2008. The influenza virus resource at the National Center for Biotechnology Information. *Journal of Virology* 82, no. 2 (January): 596–601.

Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M. 2002. An antibody that prevents the hemagglutinin low pH fusogenic transition. *Virology* 294, no. 1 (March 1): 70–74.

Barr IG, Jelley LL. 2012. The coming era of quadrivalent human influenza vaccines: who will benefit? *Drugs* 72, no. 17 (December 3): 2177–2185.

Bedford T, Rambaut A, Pascual M. 2012. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biology* 10: 38.

Bennett J, Lanning S. 2007. The Netflix Prize. *Proceedings of KDD Cup and Workshop 2007* (August 12).

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2008. GenBank. *Nucleic Acids Research* 36, no. Database issue (January): D25–30.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, et al. 2002. The Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography* 58, no. Pt 6 No 1 (June): 899–907.

Blumenthal R, Sarkar DP, Durell S, Howard DE, Morris SJ. 1996. Dilation of the influenza hemagglutinin fusion pore revealed by the kinetics of individual cell-cell fusion events. *The Journal of cell biology* 135, no. 1 (October): 63–71.

Borg I, Groenen P. 1997. *Modern Multidimensional Scaling*. Springer Series in Statistics. Berlin: Springer Press.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution* 16, no. 11 (November): 1457–65.

Cai Z, Zhang T, Wan X-F. 2010. A computational framework for influenza antigenic cartography. *PLoS Computational Biology* 6, no. 10: e1000949.

Carrat F, Flahault A. 2007. Influenza vaccine: The challenge of antigenic drift. *Vaccine* 25, no. 39-40 (September): 6852–6862.

Carter NJ, Curran MP. 2011. Live attenuated influenza vaccine (FluMist®; Fluenz^{TM}): a review of its use in the prevention of seasonal influenza in children and adults. *Drugs* 71, no. 12 (August 20): 1591–1622.

Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* 31, no. 2 Pt 1 (December): 417–27.

Chen J, Sun Y. 2011. Variation in the analysis of positively selected sites using nonsynonymous/synonymous rate ratios: an example using influenza virus. *PloS one* 6, no. 5: e19996.

Chiu C, Wrammert J, Li G-M, McCausland M, Wilson PC, Ahmed R. 2013. Cross-reactive humoral responses to influenza and their implications for a universal vaccine. *Annals of the New York Academy of Sciences* (February 13).

Clarke S, Staudt L, Kavaler J, Schwartz D, Gerhard W, Weigert M. 1990. V region gene usage and somatic mutation in the primary and secondary responses to influenza virus hemagglutinin. *J Immunol* 144, no. 7 (April 1): 2795–2801.

Clementi N, De Marco D, Mancini N, Solforosi L, Moreno GJ, Gubareva LV, Mishin V, et al. 2011. A Human Monoclonal Antibody with Neutralizing Activity against Highly Divergent Influenza Subtypes. *PloS One* 6, no. 12: e28001.

Clifford H, Wessely F, Pendurthi S, Emes RD. 2011. Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in genetics* 2: 88.

Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A, Jarrossay D, Vachieri SG, et al. 2011. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2

influenza A hemagglutinins. *Science (New York, N.Y.)* 333, no. 6044 (August 12): 850–856.

Daniels RS, Downie JC, Hay AJ, Knossow M, Skehel JJ, Wang ML, Wiley DC. 1985. Fusion mutants of the influenza virus hemagglutinin glycoprotein. *Cell* 40, no. 2 (February): 431–439.

Davies DL, Bouldin DW. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, no. 2 (April): 224 –227.

Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng P-Y, Bandaranayake D, et al. 2012. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet infectious diseases* 12, no. 9 (September): 687–695.

Dreyfus C, Laursen NS, Kwaks T, Zuijdgeest D, Khayat R, Ekiert DC, Lee JH, et al. 2012. Highly Conserved Protective Epitopes on Influenza B Viruses. *Science (New York, N.Y.)* (August 9).

Dunn JC. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* 3, no. 3: 32–57.

Edelstein L, Rosen R. 1978. Enzyme-substrate recognition. *Journal of theoretical biology* 73, no. 1 (July 6): 181–204.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32, no. 5 (March 19): 1792–1797.

Ekiert DC, Bhabha G, Elsliger M-A, Friesen RHE, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA. 2009. Antibody recognition of a highly conserved influenza virus epitope. *Science (New York, N.Y.)* 324, no. 5924 (April 10): 246–251.

Ekiert DC, Friesen RHE, Bhabha G, Kwaks T, Jongeneelen M, Yu W, Ophorst C, et al. 2011. A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science (New York, N.Y.)* 333, no. 6044 (August 12): 843–850.

Ester M, Kriegel H, Sander J, Xu X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second International Conference on Knowledge Discovery and Data Mining*, 226–231. AAAI Press.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America* 94, no. 15 (July 22): 7712–8.

Flannery B, Teukolsky S, Vetterling W. 1988. *Numerical Methods in C*. Cambridge: Cambridge University Press.

Flint SJ, Enquist LW, Racaniello VR. 2009. *Principles of Virology*. 3rd ed. ASM Press, January 1.

Gallagher P, Henneberry J, Wilson I, Sambrook J, Gething MJ. 1988. Addition of carbohydrate side chains at novel sites on influenza virus hemagglutinin can modulate the folding, transport, and activity of the molecule. *The Journal of cell biology* 107, no. 6 Pt 1 (December): 2059–2073.

Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, Vasisht N, et al. 2004. The Structure and Receptor-Binding Properties of the 1918 Influenza Hemagglutinin. *Science* (February 5): 1093155.

Gershoni JM, Roitburd-Berman A, Siman-Tov DD, Tarnovitski Freund N, Weiss Y. 2007. Epitope mapping: the first step in developing epitope-based vaccines. *BioDrugs: Clinical Immunotherapeutics, Biopharmaceuticals and Gene Therapy* 21, no. 3: 145–156.

Grafahrend-Belau E, Schreiber F, Heiner M, Sackmann A, Junker BH, Grunwald S, Speer A, Winder K, Koch I. 2008. Modularization of biochemical networks based on classification of Petri net t-invariants. *BMC Bioinformatics* 9 (February 8): 90.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52, no. 5 (October): 696–704.

Gupta R, Jung E, Brunak S. 2004. Prediction of N-glycosylation sites in human proteins. *In Preparation*.

Gupta V, Earl DJ, Deem MW. 2006. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine* 24, no. 18 (May): 3881–3888.

Hanson RM. 2010. Jmol – a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography* 43, no. 5 (September 1): 1250–1260.

Harris AK, Meyerson JR, Matsuoka Y, Kuybeda O, Moran A, Bliss D, Das SR, et al. 2013. Structure and accessibility of HA trimers on intact 2009 H1N1 pandemic influenza virus to stem region-specific neutralizing antibodies. *Proceedings of the National Academy of Sciences of the United States of America* (March 4).

Herfst S, Schrauwen EJA, Linster M, Chutinimitkul S, de Wit E, Munster VJ, Sorrell EM, et al. 2012. Airborne transmission of influenza A/H5N1 virus between ferrets. *Science (New York, N.Y.)* 336, no. 6088 (June 22): 1534–1541.

Holmes EC. 2009. The Evolutionary Genetics of Emerging Viruses. *Annual Review of Ecology, Evolution, and Systematics* 40, no. 1: 353–372.

Hu W, Chen A, Miao Y, Xia S, Ling Z, Xu K, Wang T, et al. 2013. Fully human broadly neutralizing monoclonal antibodies against influenza A viruses generated from the memory B cells of a 2009 pandemic H1N1 influenza vaccine recipient. *Virology* 435, no. 2 (January 20): 320–328.

Huang J-W, Yang J-M. 2011. Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. *BMC Bioinformatics* 12 Suppl 1: S31.

Huber PJ. 2005. *Robust Statistics*. Wiley Series in Probability and Statistics. January 28.

Imai M, Watanabe T, Hatta M, Das SC, Ozawa M, Shinya K, Zhong G, et al. 2012. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* (May 2).

Ina Y, Gojobori T. 1994. Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 18 (August 30): 8388–8392.

Jain AK. 2010. Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* 31, no. 8 (June): 651–666.

Jin H, Zhou H, Liu H, Chan W, Adhikary L, Mahmood K, Lee M-S, Kemble G. 2005. Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/Panama/2007/99. *Virology* 336, no. 1 (May 25): 113–119.

Johansson BE, Brett IC. 2008. Recombinant influenza B virus HA and NA antigens administered in equivalent amounts are immunogenically equivalent and induce equivalent homotypic and broader heterovariant protection in mice than conventional and live influenza vaccines. *Human vaccines* 4, no. 6 (December): 420–424.

Johnson NPAS, Mueller J. 2002. Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the History of Medicine* 76, no. 1: 105–15.

De Jong JC, de Ronde-Verloop FM, Veenendaal-van Herk TM, Weijers TF, Bijlsma K, Osterhaus AD. 1988. Antigenic heterogeneity within influenza A (H3N2) virus strains. *Bulletin of the World Health Organization* 66, no. 1: 47–55.

Jörg S, Ester M, Kriegel H, Xu X. 1998. *Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications*. Vol. 2. The Netherlands: Kluwer Academic Publishers.

Kashyap AK, Steel J, Rubrum A, Estelles A, Briante R, Ilyushina NA, Xu L, et al. 2010. Protection from the 2009 H1N1 pandemic influenza by an antibody from combinatorial survivor-based libraries. *PLoS pathogens* 6, no. 7: e1000990.

Kavaler J, Caton AJ, Staudt LM, Schwartz D, Gerhard W. 1990. A set of closely related antibodies dominates the primary antibody response to the antigenic site CB of the A/PR/8/34 influenza virus hemagglutinin. *Journal of immunology (Baltimore, Md.: 1950)* 145, no. 7 (October 1): 2312–2321.

Klenk HD, Garten W, Matrosovich M. 2011. Molecular mechanisms of interspecies transmission and pathogenicity of influenza viruses: Lessons from the 2009 pandemic. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* (January 13).

Kosakovsky Pond SL, Frost SDW. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England)* 21, no. 10 (May 15): 2531–2533.

Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, Brown AJL. 2006. Adaptation to Different Human Populations by HIV-1 Revealed by Codon-Based Analyses. *PLoS Computational Biology* 2, no. 6 (June 23): e62.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)* 21, no. 5 (March 1): 676–679.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Molecular Biology and Evolution* 25, no. 9 (September): 1809–24.

Kostolanský F, Varecková E, Betáková T, Mucha V, Russ G, Wharton SA. 2000. The strong positive correlation between effective affinity and infectivity neutralization of highly cross-reactive monoclonal antibody IIB4, which recognizes antigenic site B on influenza A virus haemagglutinin. *The Journal of General Virology* 81, no. Pt 7 (July): 1727–1735.

Krammer F, Pica N, Hai R, Tan GS, Palese P. 2012. Hemagglutinin Stalk-Reactive Antibodies Are Boosted following Sequential Infection with Seasonal and Pandemic H1N1 Influenza Virus in Mice. *Journal of virology* 86, no. 19 (October): 10302–10307.

Krause JC, Tumpey TM, Huffman CJ, McGraw PA, Pearce MB, Tsibane T, Hai R, Basler CF, Crowe JE Jr. 2010. Naturally occurring human monoclonal antibodies neutralize both 1918 and 2009 pandemic influenza A (H1N1) viruses. *Journal of Virology* 84, no. 6 (March): 3127–3130.

Kringelum JV, Nielsen M, Padkjær SB, Lund O. 2013. Structural analysis of B-cell epitopes in antibody:protein complexes. *Molecular immunology* 53, no. 1-2 (January): 24–34.

Krissinel E, Henrick K. 2007. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* 372, no. 3 (September 21): 774–797.

Kubota-Koketsu R, Mizuta H, Oshita M, Ideno S, Yunoki M, Kuhara M, Yamamoto N, Okuno Y, Ikuta K. 2009. Broad neutralizing human monoclonal antibodies against influenza virus from vaccinated healthy donors. *Biochemical and Biophysical Research Communications* 387, no. 1 (September 11): 180–185.

Kwong PD, Wilson IA. 2009. HIV-1 and influenza antibodies: seeing antigens in new ways. *Nature Immunology* 10, no. 6 (June): 573–578.

Laeeq S, Smith CA, Wagner SD, Thomas DB. 1997. Preferential selection of receptor-binding variants of influenza virus hemagglutinin by the neutralizing antibody repertoire of transgenic mice expressing a human immunoglobulin mu minigene. *Journal of Virology* 71, no. 4 (April): 2600–2605.

Lapedes A, Farber R. 2001. The geometry of shape space: application to influenza. *Journal of Theoretical Biology* 212, no. 1 (September 7): 57–69.

Laskowski RA. 2009. PDBsum new things. *Nucleic acids research* 37, no. Database issue (January): D355–359.

Laver WG, Air GM, Webster RG, Smith-Gill SJ. 1990. Epitopes on protein antigens: misconceptions and realities. *Cell* 61, no. 4 (May 18): 553–556.

Lee JT, Air GM. 2002. Contacts between influenza virus N9 neuraminidase and monoclonal antibody NC10. *Virology* 300, no. 2 (September 1): 255–268.

Lee M-S, Chen JS-E. 2004. Predicting antigenic variants of influenza A/H3N2 viruses. *Emerging Infectious Diseases* 10, no. 8 (August): 1385–90.

Lee PS, Yoshida R, Ekiert DC, Sakai N, Suzuki Y, Takada A, Wilson IA. 2012. Heterosubtypic antibody recognition of the influenza virus hemagglutinin receptor binding site enhanced by avidity. *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 42 (October 16): 17040–17045.

Lees WD. 2009. The Evolution of the Influenza Virus - Thesis for MRes Bioinformatics and Systems Biology. Birkbeck, University of London: UK. September 1.

Lees WD, Moss DS, Shepherd AJ. 2010. A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics* 26, no. 11 (June 1): 1403–1408.

---. 2011. Analysis of antigenically important residues in human influenza a virus in terms of B-cell epitopes. *Journal of Virology* 85, no. 17 (September): 8548–8555.

Liao Y-C, Lee M-S, Ko C-Y, Hsiung CA. 2008. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics* (January 10): 505–512.

Lin YP, Gregory V, Collins P, Kloess J, Wharton S, Cattle N, Lackenby A, Daniels R, Hay A. 2010. Neuraminidase Receptor Binding Variants of Human Influenza A(H3N2) Viruses Resulting from Substitution of Aspartic Acid 151 in the Catalytic Site: a Role in Virus Attachment? 84, no. 13 (July): 6769–6781.

Lin YP, Xiong X, Wharton SA, Martin SR, Coombs PJ, Vachieri SG, Christodoulou E, et al. 2012. Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America* (December 10).

Lingwood D, McTamney PM, Yassine HM, Whittle JRR, Guo X, Boyington JC, Wei C-J, Nabel GJ. 2012. Structural and genetic basis for development of broadly neutralizing influenza antibodies. *Nature* 489, no. 7417 (September 27): 566–570.

Lovmar L, Ahlford A, Jonsson M, Syvänen A-C. 2005. Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics* 6 (March 10): 35.

MacQueen JB. 1967. Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, 281–297. University of California Press.

Marsh GA, Rabadán R, Levine AJ, Palese P. 2008. Highly conserved regions of influenza a virus polymerase gene segments are critical for efficient viral RNA packaging. *Journal of virology* 82, no. 5 (March): 2295–2304.

Martínez-Sobrido L, Cadagan R, Steel J, Basler CF, Palese P, Moran TM, García-Sastre A. 2010. Hemagglutinin-pseudotyped green fluorescent protein-expressing influenza viruses for the detection of influenza virus neutralizing antibodies. *Journal of Virology* 84, no. 4 (February): 2157–2163.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Et Biophysica Acta* 405, no. 2 (October 20): 442–451.

Maurer-Stroh S, Lee RTC, Eisenhaber F, Cui L, Phuah SP, Lin RT. 2010. A new common mutation in the hemagglutinin of the 2009 (H1N1) influenza A virus. *PLoS Currents* 2: RRN1162.

McCauley J, Daniels R, Lin YP, Zheng X, Gregory V, Whittaker L, Cattle N, Halai C, Cross K. 2012. Report prepared for the WHO annual consultation on the composition of influenza vaccine for the Northern Hemisphere. February 20.

McDonald JC, Andrews BE. 1955. Diagnostic Methods in an Influenza Vaccine Trial. *British Medical Journal* 2, no. 4950 (November 19): 1232–1235.

Medeiros R, Escriou N, Naffakh N, Manuguerra JC, van der Werf S. 2001. Hemagglutinin residues of recent human A(H3N2) influenza viruses that contribute to the inability to agglutinate chicken erythrocytes. *Virology* 289, no. 1 (October 10): 74–85.

Meek K, Johansson B, Schulman J, Bona C, Capra JD. 1989. Nucleotide changes in sequential variants of influenza virus hemagglutinin genes and molecular structures of corresponding monoclonal antibodies specific for each variant. *Proceedings of the National Academy of Sciences of the United States of America* 86, no. 12 (June): 4664–4668.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of molecular evolution* 16, no. 1 (September): 23–36.

Morrissey B, Downard KM. 2006. A proteomics approach to survey the antigenicity of the influenza virus by mass spectrometry. *Proteomics* 6, no. 7 (April): 2034–2041.

---. 2008. Kinetics of antigen-antibody interactions employing a MALDI mass spectrometry immunoassay. *Analytical Chemistry* 80, no. 20 (October 15): 7720–7726.

Morrissey B, Streamer M, Downard KM. 2007. Antigenic characterisation of H3N2 subtypes of the influenza virus by mass spectrometry. *Journal of Virological Methods* 145, no. 2 (November): 106–114.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS genetics* 8, no. 7: e1002764.

Nabel GJ, Fauci AS. 2010. Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine. *Nature Medicine* 16, no. 12 (December): 1389–1391.

Nakajima S, Nakajima K, Nobusawa E, Zhao J, Tanaka S, Fukuzawa K. 2007. Comparison of epitope structures of H3HAs through protein modeling of influenza A virus hemagglutinin: mechanism for selection of antigenic variants in the presence of a monoclonal antibody. *Microbiology and Immunology* 51, no. 12: 1179–1187.

Nakajima S, Nobusawa E, Nakajima K. 2000. Variation in response among individuals to antigenic sites on the HA protein of human influenza virus may be responsible for the emergence of drift strains in the human population. *Virology* 274, no. 1 (August 15): 220–231.

Natali A, Oxford JS, Schild GC. 1981. Frequency of naturally occurring antibody to influenza virus antigenic variants selected in vitro with monoclonal antibody. *The Journal of Hygiene* 87, no. 2 (October): 185–190.

National Institute of Allergy and Infectious Diseases. 2013. Microbial Genome Sequencing Centers (MSC), Influenza Genome Project, Overview. URL: http://www.niaid.nih.gov/LabsAndResources/resources/dmid/gsc/Influenza/Pages/overview.aspx . Accessed 2013-05-06.

Nayak B, Kumar S, DiNapoli JM, Paldurai A, Perez DR, Collins PL, Samal SK. 2010. Contributions of the avian influenza virus HA, NA, and M2 surface proteins to the induction of neutralizing antibodies and protective immunity. *Journal of virology* 84, no. 5 (March): 2408–2420.

Ndifon W. 2011. New methods for analyzing serological data with applications to influenza surveillance. *Influenza and Other Respiratory Viruses* 5, no. 3 (May): 206–212.

Noah DL, Hill H, Hines D, White EL, Wolff MC. 2009. Qualification of the hemagglutination inhibition assay in support of pandemic influenza vaccine licensure. *Clinical and Vaccine Immunology: CVI* 16, no. 4 (April): 558–566.

Novotny J. 1991. Protein antigenicity: a thermodynamic approach. *Molecular immunology* 28, no. 3 (March): 201–207.

Okada J, Ohshima N, Kubota-Koketsu R, Iba Y, Ota S, Takase W, Yoshikawa T, et al. 2010. Localization of epitopes recognized by monoclonal antibodies that neutralized the H3N2 influenza viruses in man. *The Journal of General Virology* (November 10).

Okuno Y, Isegawa Y, Sasao F, Ueda S. 1993. A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. *Journal of Virology* 67, no. 5 (May): 2552–2558.

Osterholm M, Kelley N, Manske J, Ballering K, Leighton T, Moore K. 2012. The Compelling Need for Game-Changing Influenza Vaccines: An Analysis of the Influenza Vaccine Enterprise and Recommendations for the Future. Centre for Infectious Disease Research and Policy, University of Minnesota, October 15.

Osterholm MT, Kelley NS, Sommer A, Belongia EA. 2012. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *The Lancet infectious diseases* 12, no. 1 (January): 36–44.

Palese P, Shaw M, Knipe D, Howley P. 2007. Orthomyxoviridae - The Viruses and their Replication. In *Fields Virology*, 2: 5th ed. Philadelphia: Lippincott, Williams and Wilkins.

Palese P, Wang TT. 2011. Why do influenza virus subtypes die out? A hypothesis. *mBio* 2, no. 5.

Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, et al. 2007. Identification of the Genetic Basis for Complex Disorders by Use of Pooling-Based Genomewide Single-Nucleotide–Polymorphism Association Studies. *American Journal of Human Genetics* 80, no. 1 (January): 126–139.

Perelson AS, Oster GF. 1979. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of theoretical biology* 81, no. 4 (December 21): 645–670.

Pica N, Hai R, Krammer F, Wang TT, Maamary J, Eggink D, Tan GS, et al. 2012. Hemagglutinin stalk antibodies elicited by the 2009 pandemic influenza virus as a mechanism for the extinction of seasonal H1N1 viruses. *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 7 (February 14): 2573–2578.

Pica N, Palese P. 2013. Toward a universal influenza virus vaccine: prospects and challenges. *Annual review of medicine* 64 (January 14): 189–202.

Plotkin JB, Dushoff J. 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America* 100, no. 12 (June 10): 7152–7157.

Rand WM. 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* 66, no. 336: 846–850.

Reed ML, Bridges OA, Seiler P, Kim J-K, Yen H-L, Salomon R, Govorkova EA, Webster RG, Russell CJ. 2010. The pH of activation of the hemagglutinin protein regulates H5N1 influenza virus pathogenicity and transmissibility in ducks. *Journal of virology* 84, no. 3 (February): 1527–1535.

Reed ML, Yen H-L, DuBois RM, Bridges OA, Salomon R, Webster RG, Russell CJ. 2009. Amino acid residues in the fusion peptide pocket regulate the pH of activation of the H5N1 influenza virus hemagglutinin protein. *Journal of virology* 83, no. 8 (April): 3568–3580.

Rehermann B. 2009. Hepatitis C virus versus innate and adaptive immune responses: a tale of coevolution and coexistence. *The Journal of Clinical Investigation* 119, no. 7 (July 1): 1745–1754.

Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, no. 0 (November): 53–65.

Rubinstein ND, Mayrose I, Halperin D, Yekutieli D, Gershoni JM, Pupko T. 2008. Computational characterization of B-cell epitopes. *Molecular Immunology* 45, no. 12 (July): 3477–3489.

Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, et al. 2008. The global circulation of seasonal influenza A (H3N2) viruses. *Science (New York, N.Y.)* 320, no. 5874 (April 18): 340–6.

Russell CJ. 2011. Stalking influenza diversity with a universal antibody. *The New England Journal of Medicine* 365, no. 16 (October 20): 1541–1542.

Ryan-Poirier KA, Kawaoka Y. 1991. Distinct glycoprotein inhibitors of influenza A virus in different animal sera. *Journal of virology* 65, no. 1 (January): 389–395.

Sauter NK, Hanson JE, Glick GD, Brown JH, Crowther RL, Park SJ, Skehel JJ, Wiley DC. 1992. Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. *Biochemistry* 31, no. 40 (October 13): 9609–9621.

Schild GC, Henry-Aymard M, Pereira MS, Chakraverty P, Dowdle W, Coleman M, Chang WK. 1973. Antigenic variation in current human type A influenza viruses: antigenic characteristics of the variants and their geographic distribution. *Bulletin of the World Health Organization* 48, no. 3: 269–278.

Schmeisser F, Friedman R, Besho J, Lugovtsev V, Soto J, Wang W, Weiss C, et al. 2012. Neutralizing and protective epitopes of the 2009 pandemic influenza H1N1 hemagglutinin. *Influenza and other respiratory viruses* (November 5).

Schwahn AB, Downard KM. 2009. Antigenicity of a type A influenza virus through comparison of hemagglutination inhibition and mass spectrometry immunoassays. *Journal of Immunoassay & Immunochemistry* 30, no. 3: 245–261.

Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, et al. 2007. Anton, a special-purpose machine for molecular dynamics simulation. In *Proceedings of the 34th annual international symposium on Computer architecture*, 1–12. ISCA '07. New York, NY, USA: ACM.

Shih AC-C, Hsiao T-C, Ho M-S, Li W-H. 2007. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National*

Academy of Sciences of the United States of America 104, no. 15 (April 10): 6283–6288.

Singleton J, Santibanez T, Ding H, Lu P, Euler G, Armstrong G, Bell B, Town M, Balluz L. 2010. Interim results: influenza A (H1N1) 2009 monovalent vaccination coverage --- United States, October-December 2009. *MMWR. Morbidity and mortality weekly report* 59, no. 2 (January 22): 44–48.

Sivalingam GN, Shepherd AJ. 2012. An analysis of B-cell epitope discontinuity. *Molecular immunology* 51, no. 3-4 (July): 304–309.

Skehel J. 2009. An overview of influenza haemagglutinin and neuraminidase. *Biologicals: Journal of the International Association of Biological Standardization* 37, no. 3 (June): 177–178.

Skehel JJ, Stevens DJ, Daniels RS, Douglas AR, Knossow M, Wilson IA, Wiley DC. 1984. A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America* 81, no. 6 (March): 1779–1783.

Skehel JJ, Wiley DC. 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual Review of Biochemistry* 69: 531–69.

Smith CA, Barnett BC, Thomas DB, Temoltzin-Palacios F. 1991. Structural assignment of novel and immunodominant antigenic sites in the neutralizing antibody response of CBA/Ca mice to influenza hemagglutinin. *The Journal of Experimental Medicine* 173, no. 4 (April 1): 953–959.

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM. 2004. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* 305, no. 5682 (July 16): 371–376.

Squires RB, Noronha J, Hunt V, García-Sastre A, Macken C, Baumgarth N, Suarez D, et al. 2012. Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses* (January 20).

Staneková Z, Mucha V, Sládková T, Blaškovičová H, Kostolanský F, Varečková E. 2012. Epitope specificity of anti-HA2 antibodies induced in humans during influenza infection. *Influenza and other respiratory viruses* (January 12).

Stark SE, Caton AJ. 1991. Antibodies that are specific for a single amino acid interchange in a protein epitope use structurally distinct variable regions. *The Journal of Experimental Medicine* 174, no. 3 (September 1): 613–624.

Strengell M, Ikonen N, Ziegler T, Julkunen I. 2011. Minor changes in the hemagglutinin of influenza A(H1N1)2009 virus alter its antigenic properties. *PloS One* 6, no. 10: e25848.

Sugar CA, James GM. 2003. Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association* 98, no. 463: 750–763.

Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen L, Santelli E, et al. 2009. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature Structural & Molecular Biology* 16, no. 3 (March): 265–273.

Sui J, Sheehan J, Hwang WC, Bankston LA, Burchett SK, Huang C-Y, Liddington RC, Beigel JH, Marasco WA. 2011. Wide prevalence of heterosubtypic broadly neutralizing human anti-influenza A antibodies. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 52, no. 8 (April 15): 1003–1009.

Suzuki Y. 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Molecular Biology and Evolution* 21, no. 12 (December): 2352–2359.

---. 2006. Natural selection on the influenza virus genome. *Molecular biology and evolution* 23, no. 10 (October): 1902–1911.

---. 2008. Positive selection operates continuously on hemagglutinin during evolution of H3N2 human influenza A virus. *Gene* 427, no. 1-2 (December 31): 111–6.

---. 2011. Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus. *Genes & Genetic Systems* 86, no. 5: 287–294.

---. 2013. Detection of positive selection eliminating effects of structural constraints in hemagglutinin of H3N2 human influenza A virus. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases* (February 9).

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* 16, no. 10 (October): 1315–1328.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution* 10, no. 3 (May): 512–526.

Tavaré S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. In *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, 17:57–86. Amer Mathematical Society.

Thoennes S, Li Z-N, Lee B-J, Langley WA, Skehel JJ, Russell RJ, Steinhauer DA. 2008. Analysis of residues near the fusion peptide in the influenza hemagglutinin structure for roles in triggering membrane fusion. *Virology* 370, no. 2 (January 20): 403–414.

Throsby M, van den Brink E, Jongeneelen M, Poon LLM, Alard P, Cornelissen L, Bakker A, et al. 2008. Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PloS One* 3, no. 12: e3942.

Tong S, Li Y, Rivailler P, Conrardy C, Castillo DAA, Chen L-M, Recuenco S, et al. 2012. A distinct lineage of influenza A virus from bats. *Proceedings of the National Academy of Sciences* 109, no. 11 (March 13): 4269–4274.

Tsibane T, Ekiert DC, Krause JC, Martinez O, Crowe JE Jr, Wilson IA, Basler CF. 2012. Influenza Human Monoclonal Antibody 1F1 Interacts with Three Major Antigenic Sites and Residues Mediating Human Receptor Specificity in H1N1 Viruses. *PLoS pathogens* 8, no. 12 (December): e1003067.

Tusche C, Steinbrück L, McHardy AC. 2012. Detecting patches of protein sites of influenza A viruses under positive selection. *Molecular Biology and Evolution* (March 16).

Underwood PA. 1984. An antigenic map of the haemagglutinin of the influenza Hong Kong subtype (H3N2), constructed using mouse monoclonal antibodies. *Molecular Immunology* 21, no. 7 (July): 663–671.

Vanlandschoot P, Beirnaert E, Barrère B, Calder L, Millar B, Wharton S, Jou WM, Fiers W. 1998. An antibody which binds to the membrane-proximal end of influenza virus haemagglutinin (H3 subtype) inhibits the low-pH-induced conformational change and cell-cell fusion but does not neutralize virus. *The Journal of General Virology* 79 (Pt 7) (July): 1781–1791.

Varecková E, Mucha V, Wharton SA, Kostolanský F. 2003. Inhibition of fusion activity of influenza A haemagglutinin mediated by HA2-specific monoclonal antibodies. *Archives of Virology* 148, no. 3 (March): 469–486.

Venkatramani L, Bochkareva E, Lee JT, Gulati U, Graeme Laver W, Bochkarev A, Air GM. 2006. An epidemiologically significant epitope of a 1998 human influenza virus neuraminidase forms a highly hydrated interface in the NA-antibody complex. *Journal of molecular biology* 356, no. 3 (February 24): 651–663.

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. 2010. The immune epitope database 2.0. *Nucleic Acids Research* 38, no. Database issue (January): D854–862.

Wang TT, Palese P. 2009. Universal epitopes of influenza virus hemagglutinins? *Nature Structural & Molecular Biology* 16, no. 3 (March): 233–234.

Wang TT, Tan GS, Hai R, Pica N, Petersen E, Moran TM, Palese P. 2010. Broadly protective monoclonal antibodies against H3 influenza viruses following sequential immunization with different hemagglutinins. *PLoS Pathogens* 6, no. 2: e1000796.

Wang W, Anderson CM, De Feo CJ, Zhuang M, Yang H, Vassell R, Xie H, Ye Z, Scott D, Weiss CD. 2011. Cross-neutralizing antibodies to pandemic 2009 H1N1 and recent seasonal H1N1 influenza A strains influenced by a mutation in hemagglutinin subunit 2. *PLoS pathogens* 7, no. 6 (June): e1002081.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)* 25, no. 9 (May 1): 1189–1191.

Webster RG, Hinshaw VS, Laver WG. 1982. Selection and analysis of antigenic variants of the neuraminidase of N2 influenza viruses with monoclonal antibodies. *Virology* 117, no. 1 (February): 93–104.

Wei C-J, Boyington JC, McTamney PM, Kong W-P, Pearce MB, Xu L, Andersen H, et al. 2010. Induction of broadly neutralizing H1N1 influenza antibodies by vaccination. *Science (New York, N.Y.)* 329, no. 5995 (August 27): 1060–1064.

Wiley DC, Skehel JJ. 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Review of Biochemistry* 56: 365–94.

Wiley DC, Wilson IA, Skehel JJ. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 289, no. 5796 (January 29): 373–8.

Wilson IA, Cox N. 1990. Structural basis of immune recognition of influenza virus hemagglutinin. *Annual review of immunology* 8: 737–771.

Wilson IA, Skehel JJ, Wiley DC. 1981. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. *Nature* 289, no. 5796 (January 29): 366–73.

Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, Anderson EC, Barclay WS, Digard P. 2009. A Complicated Message: Identification of a Novel PB1-Related Protein Translated from Influenza A Virus Segment 2 mRNA. *Journal of Virology* 83, no. 16 (August 15): 8021–8031.

Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. 2006. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct* 1: 34.

World Health Organization. 2011. *Manual for the laboratory diagnosis and virological surveillance of influenza*.

---. 2011. *WHO information for molecular diagnosis of influenza virus in humans*. August.

Wrammert J, Koutsonanos D, Li G-M, Edupuganti S, Sui J, Morrissey M, McCausland M, et al. 2011. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *The Journal of Experimental Medicine* 208, no. 1 (January 17): 181–193.

Wright PF, Neumann G, Kawaoka Y. 2007. Orthomyxoviruses. In *Fields Virology*, ed by. David Knipe and Peter Howley, 2: 5th ed. Philadelphia: Lippincott, Williams and WIlkins.

Xia Z, Huynh T, Kang S-G, Zhou R. 2012. Free-Energy Simulations Reveal that Both Hydrophobic and Polar Interactions Are Important for Influenza Hemagglutinin Antibody Binding. *Biophysical Journal* 102, no. 6 (March 21): 1453–1461.

Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE, Wilson IA. 2010. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science (New York, N.Y.)* 328, no. 5976 (April 16): 357–360.

Yamada A, Nobusawa E, Cao MS, Imanishi J, Oyama S, Abe A, Katagiri S, Kim DW, Nakajima K, Nakajima S. 1991. Epitope changes on the haemagglutinin molecule of recently isolated H1N1 influenza viruses. *The Journal of General Virology* 72 (Pt 1) (January): 97–102.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of molecular evolution* 39, no. 3 (September): 306–314.

---. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution* 51, no. 5 (November): 423–32.

---. 2006. *Computational Molecular Evolution*. Oxford University Press, December 7.

Yewdell JW. 2011. Viva la Revolución: Rethinking Influenza A Virus Antigenic Drift. *Current opinion in virology* 1, no. 3 (September 1): 177–183.

Yewdell JW, Caton AJ, Gerhard W. 1986. Selection of influenza A virus adsorptive mutants by growth in the presence of a mixture of monoclonal antihemagglutinin antibodies. *Journal of Virology* 57, no. 2 (February): 623–628.

Yu X, Tsibane T, McGraw PA, House FS, Keefer CJ, Hicar MD, Tumpey TM, et al. 2008. Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* 455, no. 7212 (September 25): 532–536.

Zhou R, Das P, Royyuru A. 2008. Single Mutation Induced H3N2 Hemagglutinin Antibody Neutralization: A Free Energy Perturbation Study. *The Journal of Physical Chemistry. B* (November 1).